IMAGE CORRESPONDENCES FROM PERCEIVED MOTION

by

Richard Kirby

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Computing

School of Computing

The University of Utah

May 2017

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Richard Kirby**

has been approved by the following supervisory committee members:

Ross Whitaker	, Chair	**3/6/2017** Date Approved
Guido Gerig	, Member	**2/28/2017** Date Approved
John Hollerbach	, Member	**2/17/2017** Date Approved
Thomas Henderson	, Member	**2/17/2017** Date Approved
Srikumar Ramalingam	, Member	**2/17/2017** Date Approved

and by **Ross Whitaker** , Chair/Dean of

the Department/College/School of **Computing**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

3D reconstruction from image pairs relies on finding corresponding points between images and using the corresponding points to estimate a dense disparity map. Today's correspondence-finding algorithms primarily use image features or pixel intensities common between image pairs. Some 3D computer vision applications, however, don't produce the desired results using correspondences derived from image features or pixel intensities. Two examples are the multimodal camera rig and the center region of a coaxial camera rig. Additionally, traditional stereo correspondence-finding techniques which use image features or pixel intensities sometimes produce inaccurate results. This thesis presents a novel image correspondence-finding technique that aligns pairs of image sequences using the optical flow fields. The optical flow fields provide information about the structure and motion of the scene which is not available in still images, but which can be used to align images taken from different camera positions.

The method applies to applications where there is inherent motion between the camera rig and the scene and where the scene has enough visual texture to produce optical flow. We apply the technique to a traditional binocular stereo rig consisting of an RGB/IR camera pair and to a coaxial camera rig. We present results for synthetic flow fields and for real images sequences with accuracy metrics and reconstructed depth maps.

To Mattie.

"If we knew what it was we were doing, it would not be called research, would it?"

Albert Einstein

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

This dissertation would not be possible without the contribution and support of many people. I especially want to thank my committee and in particular my advisor Ross Whitaker. Ross has an extraordinary way of challenging my thinking in a way that increases my motivation. His guidance instilled an academic discipline in me that was key to working through the challenges of this dissertation. His fondness for elegant simplicity made the equations and derivations far more readable than they would have been. Guido Gerig provided key insights in computer vision. It was one of his questions on the qualification exam that lead to the dual focal length camera solution that solved the issues with frontal planar surfaces. Jur van den Berg, as my first advisor, was always available to work through ideas and connected my interests in computer vision to robotics. Jur has a talent for asking the right questions. Tom Henderson was always accessible to bounce ideas off of and was key in finding flaws in my mathematical logic. John Hollerbach provided important guidance at several junctures in my graduate studies that kept me moving in the right direction. Without his guidance, it's unlikely I would have reached this point.

To Mattie, my wife and soul mate. Mattie keeps me grounded, provides constant inspiration, and creates the environment that promotes creative thinking and hard work. Her talent for listening helps more than she realizes.

CHAPTER 1

INTRODUCTION

The goal of this dissertation is to develop a methodology to find correspondences between optical flow fields derived from pairs of image sequences (Figure 1.1). These correspondences, along with the mathematical relationship between the flow fields at corresponding pixel locations, is used to estimate both the dense depth map and the scene flow (dense three-dimensional motion field). This estimation is done without directly using intercamera image features or pixel intensities, which permits image alignment, dense depth map estimation, and scene flow estimation in image pairs where traditional image-feature or pixel-intensity-based methods may be inadequate. The method is tested on image sequences from two

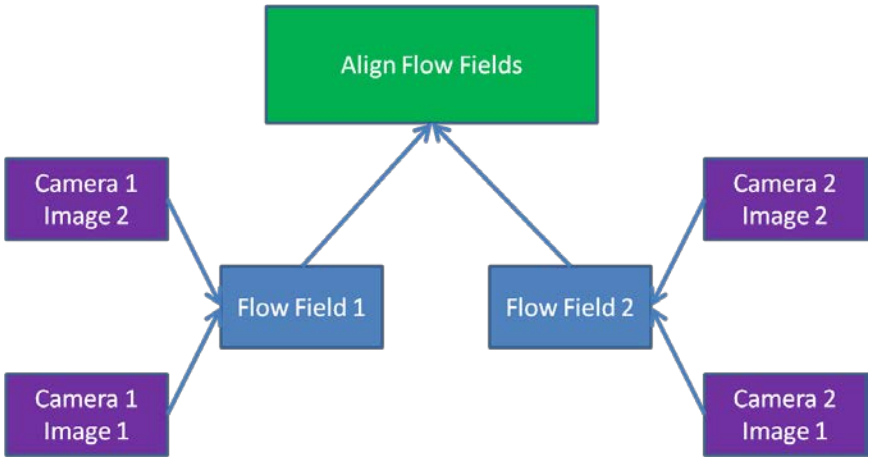


Figure 1.1. Methodology developed in this dissertation.

multicamera rigs, a coaxial camera rig with collinear optical axes consisting of two color (RGB) cameras (Figure 1.2a), and a multimodal camera rig with parallel optical axes consisting of an RGB camera and an infrared (IR) camera (Figure 1.2b).

Optical flow fields contain information about the scene that is not available in still image pairs, namely a representation of the scene motion encoded by the scene shape. Unlike traditional image correspondences derived from intercamera image features or pixel intensities, in all but the simplest cases, optical flow fields taken from different camera positions will be different. In this dissertation, the mathematical relationship between optical flow fields taken from different camera positions is derived with minimal assumptions about the structure of the scene. Correspondences are found between flow fields taken from different viewpoints by using an energy-minimization approach based on this mathematical relationship. This process produces dense optical flow field matches in image sequence pairs that permit image alignment based on the perceived

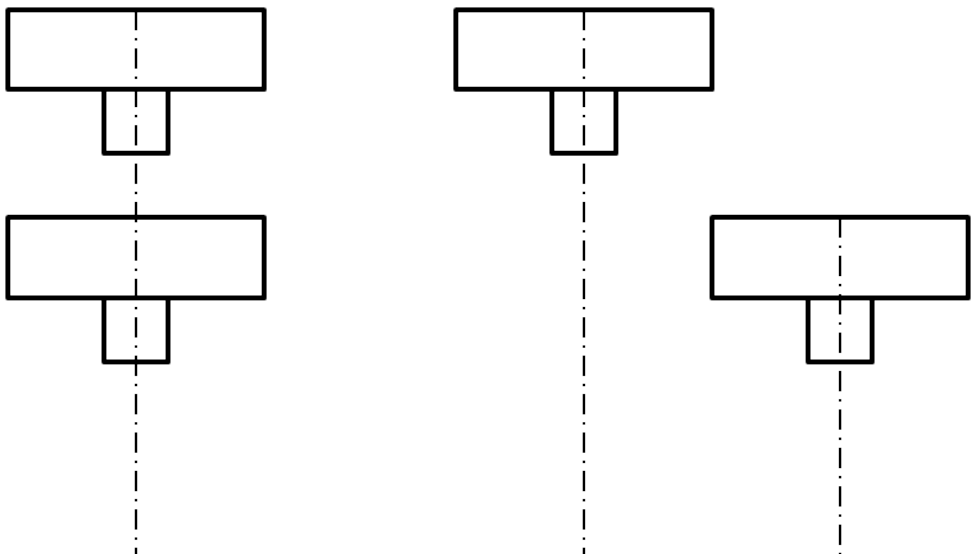


Figure 1.2. The two types of camera rigs used in this dissertation. (a) Coaxial (RGB/RGB) camera rig. (b) Multimodal (RGB/IR) camera rig.

motion in the images.

Optical flow field derived image correspondences have several unique characteristics that allow the finding of image correspondences in situations where intercamera image correspondences derived from pixel intensities or image features do not produce good results. First, optical flow fields are invariant to the wavelength of light being imaged as long as the images have sufficient features visible at each wavelength imaged. This invariance allows flow field alignment taken with pairs of cameras that image different wavelengths to produce dense depth maps. Second, optical flow fields obtained using cameras with different focal length optical systems produce a unique relationship between the disparity of corresponding points in the two flow fields and the optical flow reported in those corresponding pixels. The disparity allows one image to be warped into alignment with the other; however, it is the ratio of the optical flow field values at a given disparity that contains the information used to estimate depth. This unique characteristic of flow field alignment allows depth to be estimated where there is no pixel disparity, something that is impossible with pixel-intensity or feature-based correspondences.

While useful as a stand-alone correspondence finding technique where traditional intercamera image-feature- or pixel-intensity-based approaches fail, optical flow field derived correspondences also have applications when used in combination with intra-camera image-feature- and/or pixel-intensity-based approaches. The estimation of scene flow typically uses both optical flow and stereo correspondences [1]-[6]. Scene flow methods that decouple the depth estimation from the motion estimation are reported to have advantages over methods that combine the disparity estimation and motion estimation into a single framework [3]. However, the decoupled methods do not take

advantage of the additional information that comes from aligning the optical flow fields. Instead, decoupled scene flow methods find depth from stereo correspondences and use the optical flow to track points over time. Combining optical flow field derived correspondences with intercamera image-feature- or pixel-intensity-based correspondences produces a redundant set of correspondences based on different scene information (intensities and/or features vs. motion). Where the two sets of correspondences do not match, they provide insight into the error in the decoupled scene flow estimation as well as a consistency constraint in the optical flow computation. One could foreseeably use this optical flow consistency constraint to improve the estimation of the optical flow, thereby improving the overall accuracy of decoupled scene flow estimation.

1.1 Motivation

The motivation for undertaking the research presented in this dissertation originated from the need to measure movement in several situations where no existing technology was feasible. In earlier work of the author, optical flow was used as a measurement tool in several specialized devices for real-time analysis of the technique of athletes. These devices used commercially available optical flow sensors in several different multisensor configurations to estimate three-dimensional (3D) motion. Optical flow sensors produce reasonably accurate estimates of 3D velocity, but due to limitations of the sensor design, work only in very selective environments. The following section briefly describes this early work, along with the accuracies achieved and the limitations uncovered. This work suggested that generalizing the methodology to multicamera rigs could result in a system

capable of estimating dense depth maps and scene flow in situations where existing methods were not feasible.

1.2 Early Work

The first time optical flow alone was used to make depth estimations was in a system designed to provide real-time technique feedback to world-class skiers [7]-[9]. The device was called the vLink Racing Computer (Figure 1.3) and consisted of a pair of ski-mounted sensor units, each containing a commercial optical flow sensor. The design objective of the system was to measure lateral slippage (which reduces the speed of a ski racer) and deliver real-time audible feedback to the athlete. The feedback was proportional to the lateral slippage of the skier, enabling the skier to correct technique

Figure 1.3. vLink racing computer.

errors in real-time that caused excessive lateral slippage. Two sensors were required so that at any time at least one of the sensors, the one on the edge of the ski being used to make the turn, was in contact with the snow surface. During the development process it was recognized that the information from the sensor not in contact with the snow also produced velocity information that was scaled by its distance from the snow surface (Figure 1.4). The ratio of the two velocities from the two different optical flow measurements produced an estimate of the angle of the ski relative to the plane of the snow. This edge-angle estimate became a valuable component in analyzing skier technique. It also demonstrated that the ratio of the perceived velocity could be used to estimate depth.

In addition to the ability to estimate depth, the work with the vLink revealed that under the right conditions, optical flow could be used to accurately measure velocity relative to visually textured frontal planar surfaces. In controlled experiments a standard deviation of 9 mm over a 100 m test track during 18 consecutive trials was measured,

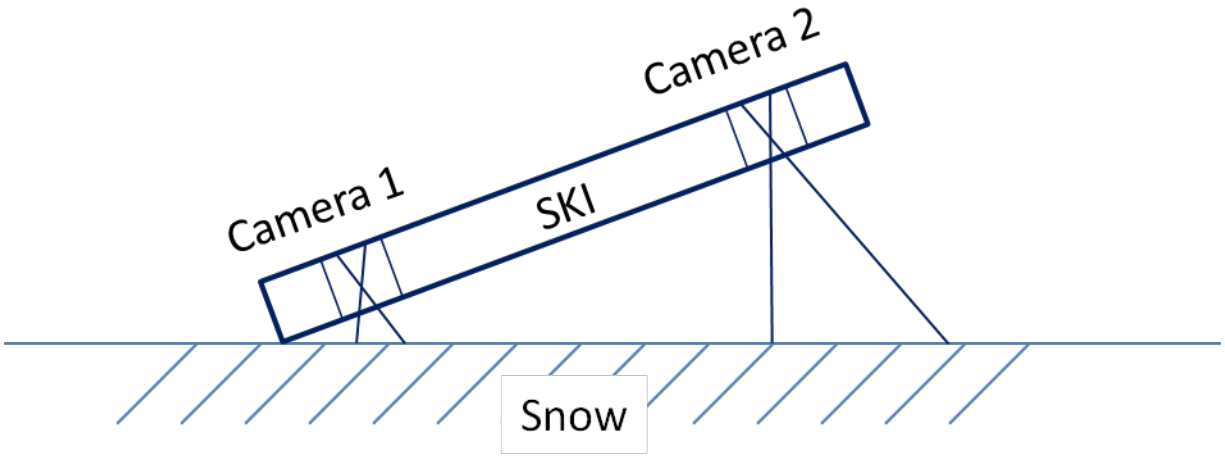


Figure 1.4. Ski edge angle from optical flow triangulation.

which equates to a standard error of 0.009%. Not only is the accuracy high relative to other technologies used to measure skier velocity (GPS, IMUs, and passive marker motion capture), but the velocity estimation is done in real-time, making real-time feedback possible.

Based on the insights gained during the development of the vLink training device, two other products that involve measuring velocity on snow were developed. The first was a glide test instrument used today in the development of ski waxes. It produces the most accurate velocity and distant estimates of any ski sensor system [10], and when the data are fused with data from Hall effect sensors and accelerometers, the instrument is capable of estimating the speed-dependent friction coefficient between the ski base and the snow surface [11]. The second device is an electronic avalanche probe [12] that measures snow-pack density using a force sensor at precise depth intervals. The depth intervals are computed from optical flow data, and the result is a snow-pack density profile used to predict avalanches.

A system similar to the vLink, but intended for analyzing a golf swing, was designed for TaylorMade Golf, using three optical flow sensors (Figure 1.5). Unlike skis, however, a golf club has additional degrees of freedom that complicate the estimation of 3D motion. For example, the difference in perceived velocity between two sensors in a stereo configuration varies by depth as well as by club head rotation. This problem was resolved using a coaxial sensor arrangement [13], [14], which effectively isolates the flow field differences due to depth from other forms of motion. This coaxial sensor arrangement has a number of additional advantages over traditional binocular stereo that make it more suitable to certain embedded and space-constrained applications.

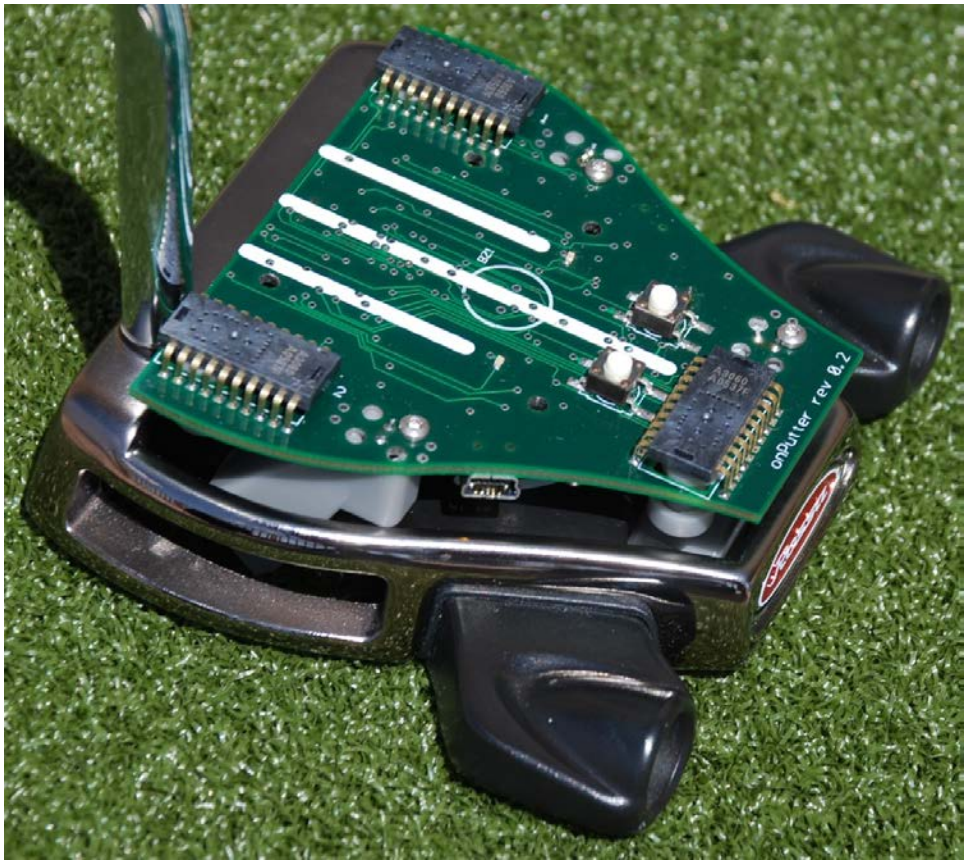


Figure 1.5. Prototype three-sensor optical flow putter.

Specifically, unlike sensors in a binocular stereo configuration, sensors in a coaxial configuration have a minimum working distance that is primarily constrained by the ability to focus the image, rather than the intercamera image overlap. Second, the baseline is not constrained by the working distance, allowing large baselines and thus higher precision with very small working distances. Third, the baseline of a coaxial sensor configuration can be parallel or perpendicular to the image plane, or even a combination of the two, which allows large baselines to be wrapped up inside the camera rig, permitting depth estimation through a narrow diameter tube. Lastly, a coaxial sensor arrangement has substantially smaller occlusions than a binocular stereo camera rig.

These advantages, however, come at a cost. As will be seen in detail in Chapter 2, depth from a coaxial camera rig cannot be estimated using intercamera image-feature- or pixel-intensity-based correspondences in the central region of the image. Additionally, the optical flow sensors used in the work described here produce a single flow value over a small field of view. This single flow value essentially takes advantage of a singularity along the optical axis that allows depth to be computed directly from the ratio of the flow computed by two sensors. This singularity does not extend to off-axis computation of optical flow, which complicates the estimate of dense depth maps and dense 3D motion fields.

This early work was intended to solve very specific problems in motion estimation, but the accuracy of the velocity estimation combined with the ability to estimate depth suggests that if the method could be generalized to work with optical flow fields acquired with multicamera rigs, it would provide a valuable tool in situations where existing methods do not produce good results. The initial motivation for this research was centered around the coaxial camera rig, but the ability to estimate dense depth maps and dense 3D motion fields without using intercamera image-feature- or pixel-intensity-based correspondences extends beyond the coaxial camera rig.

To generalize the method such that it works with optical flow fields computed from image sequences acquired by multicamera rigs, several problems need to be overcome. First, optical flow fields taken at different distances and through different focal length imaging systems need to be aligned, which requires finding optical flow field based correspondences. This problem is more difficult than finding image-feature or pixel-intensity-based correspondences because unlike image features or pixel intensities,

different camera positions produce different optical flow fields. Solving this problem requires finding the mathematical relationship between flow fields taken from different camera positions. Second, once the relationship between flow fields is found, it must be incorporated into an intercamera optical flow-field correspondence-finding formulation. Third, an efficient numerical solution to the flow-field correspondence-finding formulation must be developed. Fourth, frontal planar surfaces, which are common in many scenes, produce constant flow, which for a method that aligns optical flow fields is equivalent to a featureless region for methods that align images based on image feature or pixel intensities. A way of aligning constant flow regions is required for the technique to have broad applicability.

1.3 Contributions

The contributions of this dissertation include:

1) The derivation of the relationship between two optical flow fields taken from different camera positions for two multicamera geometry types, a coaxial camera rig and a multimodal stereo camera rig. These derivations provide a model for deriving the relationships between flow fields for other multicamera configurations.
2) The derivation of an energy-minimization functional using the mathematical relationship between flow fields to align them in such a way that results in alignment of the underlying images.
3) Two numerical solutions to the energy-minimization problem, one using variational methods and the second using graph cuts.

4) Camera rig and imaging optics considerations required to align frontal planar regions—a degenerate case for stereo camera rigs using optical flow for image alignment.

5) The derivation of the equations required to convert the aligned optical flow fields into dense depth maps and scene flow.

6) Results with accuracy metrics for the technique with the two types of camera rigs described above on three real-world scenes, including comparisons with the state-of-the-art multimodal and structure from motion (SfM) algorithms on the same three real-world scenes.

7) A discussion of how the methodology can be applied to solve two real-world problems. The first is 3D reconstruction via an endoscope with a single opening for images entering the optical system, and the second is 3D reconstruction from image sequences taken with a multimodal RGB/IR stereo camera rig common in many surveillance applications.

1.4 Overview

Chapter 2 describes the coaxial camera rig; presents the literature on depth from zooming—the predecessor of the coaxial camera; derives the relationship between the optical flow fields acquired by the two cameras; and develops the energy-minimization functional that, when solved, results in the alignment of the two flow fields. Two solutions to the energy-minimization problem are presented, a variational methods approach and an approach using graph techniques. The method is demonstrated on three real-world scenes, and accuracy metrics are presented and compared with a state-of-the-

art scaled SfM algorithm.

Chapter 3 describes the multimodal stereo camera rig and presents the literature for image alignment and depth estimation using images generated by cameras that image different light frequencies. The relationship between the optical flow fields acquired by the two cameras with different focal lengths is derived, and the energy-minimization functional that, when solved, results in the alignment of the two flow fields is developed. Both a variational methods approach and a graph-technique approach are used to solve the energy functional, and the technique is demonstrated on a number of real-world images. The accuracy of alignment metrics is presented and compared with the state-of-the-art multimodal method, which uses a combination of scale-invariant feature transform (SIFT) and edge-oriented histograms (EOH) to produce sparse points of interest matches in multimodal image pairs.

Conclusions are presented in Chapter 4 with a discussion of how the methodology can be applied to existing scene flow algorithms to add an additional constraint to the optical flow computation that could result in more accurate scene flow estimation.

CHAPTER 2

THE COAXIAL CAMERA RIG

2.1 Introduction

A coaxial camera rig consists of two cameras that image along the same optical axis using a beam splitter to create two independent optical paths (Figure 2.1). The two optical paths are imaged via two independent optical systems having different focal lengths and different distances to the scene. The different focal lengths combined with the different distances to the scene produce different magnifications of the scene onto the image sensor. The different magnifications result in a radial disparity that is a function of the depth between the camera rig and the scene. However, because the images acquired by each camera share the same optical axis, the disparity is always zero on the optical axis and very small in the center region of the image pairs, which makes it impossible to recover depth in the center region of a coaxial camera using pixel disparities.

If not for the lack of disparity in the center region, coaxial cameras would likely be more prevalent as they have a number of advantages over traditional binocular stereo camera rigs, including:

1) There are fewer and smaller occlusions as alluded to by Ma and Olsen [15].
2) The baseline can be wrapped up inside the camera in such a way that the camera rig can acquire images through a small diameter tube [16].

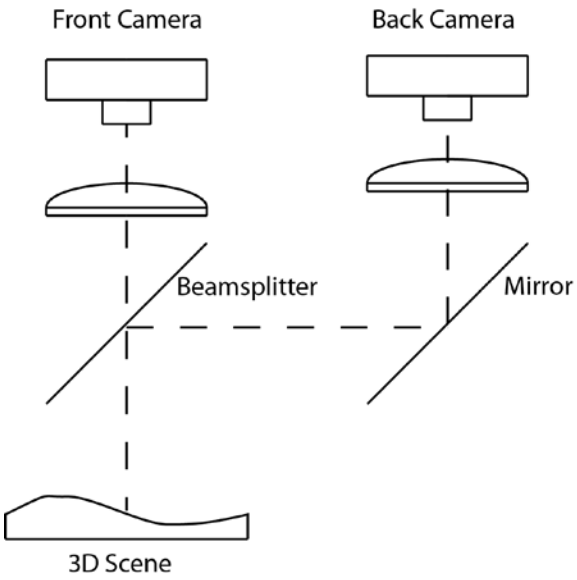


Figure 2.1. Schematic representation of a coaxial camera rig.

3) A singularity along the optical axis substantially reduces the complexity of depth computation when the scene is rigid, and relative motion consists of frontal planar translation. This singularity also provides good values for initialization in other configurations.
4) The center point produces one known correspondence value for all radial epipolar lines.
5) Unlike a binocular stereo camera rig, a coaxial camera rig has no minimum working distance other than the ability to focus.

If it were possible to estimate depth in the center region of a coaxial camera, this type of camera rig would have a myriad of uses from embedded applications (such as skiing and golf clubs), to space constrained applications that require imaging through a tube such as in a borescope or endoscope.

2.2 Related Work - Depth from Zooming

Outside the publications associated with this dissertation [13], [16], [17], there is little literature on the coaxial camera. However, estimating depth from images taken at different focal lengths by changing the zoom on a fixed camera imaging a stationary scene has been known for many years as depth from zooming. A coaxial camera rig is fundamentally a simultaneous depth from zooming setup.

Estimating dense depth maps from a depth from zooming setup was first proposed by Ma and Olsen [15] in 1990. Lavest et al. [18], [19] provide a proof for inferring 3D data from images taken at multiple focal lengths and model a revolving object. Asada et al. [20] and Baba et al. [21] present a method for doing 3D reconstruction using blur from zoom. Gao et al. [22] present a distance measurement system for mobile robots using zooming. Most recently, Zhang and Qi [23] describe a method for 3D reconstruction by finding corresponding contours (using a snake search algorithm) between images taken at different focal lengths and then using the camera geometry to estimate depth along the contours.

The primary reason researchers have investigated using a single zoom camera to do 3D reconstruction is cost. As noted above, however, depth from zooming has some additional advantages if the stationary scene constraint and unrecoverable point problem could be overcome. The coaxial camera rig combined with image correspondences derived from perceived motion overcomes these problems. First, simultaneous images taken at two different focal lengths overcome the stationary scene constraint of depth from zooming. Second, using the flow fields to align image pairs overcomes the unrecoverable point problem in the center region described by Ma and Olsen [15]. This

later advantage is due to the depth estimate being derived from the ratio of the flow fields taken at different focal lengths as opposed to the extremely small disparities found in the center region of a coaxial camera rig.

2.3 Energy Formulation

Referring to Figure 2.2, let $\bar{x}_f = (x_f, y_f)^T$ and $\bar{x}_b = (x_b, y_b)^T$ represent points in the image domain of the front and back cameras. Let $\bar{h}(\bar{x})$ be the disparity between $\bar{x}_f$ and $\bar{x}_b$ such that $\bar{x}_f$ and $\bar{x}_b - \bar{h}(\bar{x}_f)$ represent the same point $\bar{X}(\bar{x}_f) = (X, Y)$ in the scene. Let f_f and f_b be the focal lengths for the front camera and back cameras and $Z(\bar{x}_f)$ be the distance between the optical center of the front camera and a point in the scene corresponding to $\bar{x}_f$, the distance being measured along the optical axis. Let b be the distance between the optical center of the two cameras. Let $\bar{w}_f$ and $\bar{w}_b$ be the projection of the 3D motion field onto the image planes of the front and back cameras, respectively.

In Figure 2.2, the Y axis is pointing out of the page. Because the coaxial camera is symmetrical around the Z axis, equations derived for the X-Z plane are identical to those derived for the Y-Z plane.

We first derive equations for the disparity $\bar{h}(\bar{x})$. We start with the projection equations for a pinhole camera

$$\bar{x}_f = -\frac{f_f \bar{X}}{Z} \tag{2.1}$$

$$\bar{x}_b = -\frac{f_b \bar{X}}{Z+b} \tag{2.2}$$

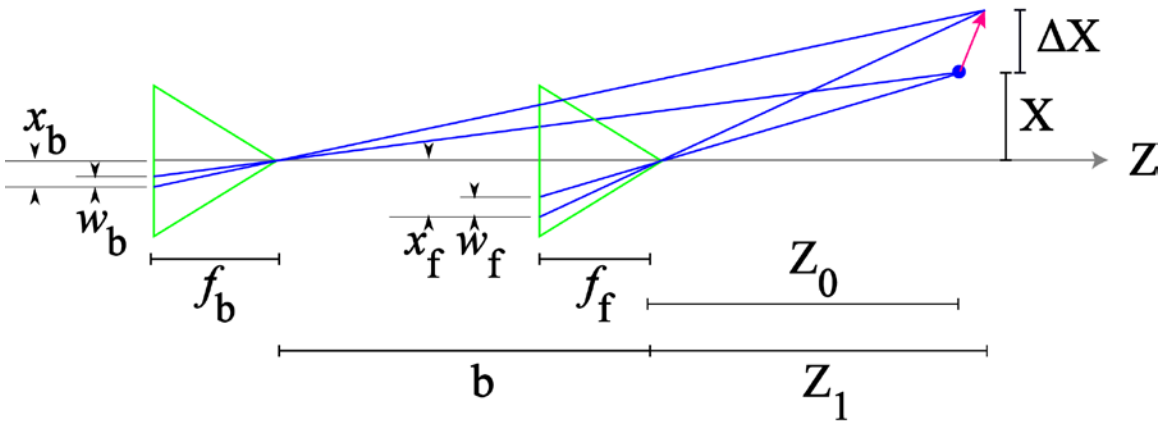


Figure 2.2. Coaxial camera rig geometry.

Solving for the disparity gives

$$\bar{x}_f - \bar{x}_b = \bar{h}(\bar{x}_f) = \frac{f_f \bar{X}}{Z} - \frac{f_b \bar{X}}{Z+b} \tag{2.3}$$

which, when reduced, results in

$$\bar{h}(\bar{x}_f) = \frac{\bar{x}_f\left(\frac{f_b}{f_f}Z - Z - b\right)}{Z+b}. \tag{2.4}$$

We next find the relationship between the optical flow perceived by the two cameras. Once again, we start with the projection equations and take the derivative with respect to time

$$\frac{dx}{dt} = w_x = -f\frac{d}{dt}\left(\frac{X}{Z}\right) \tag{2.5}$$

$$\frac{dy}{dt} = w_y = -f\frac{d}{dt}\left(\frac{Y}{Z}\right) \tag{2.6}$$

$$w_x = \frac{x\dot{Z} - f\dot{X}}{Z} \tag{2.7}$$

$$w_y = \frac{y\dot{Z} - f\dot{Y}}{Z} \tag{2.8}$$

which can be written in homogeneous coordinates as

$$\bar{P} = \begin{bmatrix} 1 & 0 & x/f & 0 \\ 0 & 1 & y/f & 0 \\ 0 & 0 & 0 & -Z/f \end{bmatrix} \tag{2.9}$$

$$\bar{w} = \begin{bmatrix} 1 & 0 & -x/f & 0 \\ 0 & 1 & -y/f & 0 \\ 0 & 0 & 0 & -Z/f \end{bmatrix} \begin{bmatrix} \dot{X} \\ \dot{Y} \\ \dot{Z} \\ 1 \end{bmatrix} = \begin{bmatrix} \dot{X} - \frac{x\dot{Z}}{f} \\ \dot{Y} - \frac{y\dot{Z}}{f} \\ -\frac{Z}{f} \end{bmatrix} = -\begin{bmatrix} \frac{f\dot{X} - x\dot{Z}}{Z} \\ \frac{f\dot{Y} - y\dot{Z}}{Z} \end{bmatrix} \tag{2.10}$$

Adding image-frame timing and adding the different baseline for the front and back cameras, (2.7) and (2.8) become

$$\bar{w}_f = \frac{\bar{x}_{f0}\dot{Z} - f_f\dot{\bar{X}}}{Z_{f1}} \tag{2.11}$$

$$\bar{w}_b = \frac{\bar{x}_{b0}\dot{Z} - f_b\dot{\bar{X}}}{Z_{b1}}. \tag{2.12}$$

Solving for $\dot{\bar{X}}$ and setting the resulting equations equal to each other gives

$$m(\bar{x}_f)\bar{w}_f(\bar{x}_f) = c(\bar{x}_f)\bar{w}_b\left(\bar{x}_f + \bar{h}(\bar{x}_f)\right) \tag{2.13}$$

where

$$m(\bar{x}_f) = \left(\frac{f_b}{f_f}\right)\left(\frac{Z(\bar{x}_f)}{(Z(\bar{x}_f)+b)}\right) \tag{2.14}$$

and

$$\bar{c}(\bar{x}_f) = \left(\frac{\bar{w}_f(\bar{x}_f)}{\left(\frac{Z_0(\bar{x}_f)+b}{Z_1(\bar{x}_f)+b}\right)\left(\frac{Z_1(\bar{x}_f)}{Z_0(\bar{x}_f)}\right)(\bar{w}_f(\bar{x}_f)+\bar{x}_f)-\bar{x}_f}\right) \tag{2.15}$$

for the front and back cameras. Equation (2.13) can be written as the energy functional

$$E_{match} = \sum_{p\in\mathcal{P}}\left[\bar{p}(\bar{x}_f)\bar{w}_f(\bar{x}_f) - \bar{c}(\bar{x}_f)\bar{w}_b\left(\bar{x}_f + \bar{h}(\bar{x}_f)\right)\right]^2. \tag{2.16}$$

Equation (2.16) has a similar form to the data term in many optical flow algorithms (see [24]). However, instead of assigning a cost penalty based on how well pixel intensities match, (2.16) assigns a cost penalty based on how well the optical flow fields match. Perfectly matched optical flow fields produce a zero cost penalty. Like optical flow and many other inverse problems, this problem is ill-posed (e.g., a unique solution may not exist). Ill-posedness is commonly overcome by using a regularization term that penalizes some kind variation. For optical flow the regularization term penalizes

variations in the flow field. The regularization term used here penalizes variations in depth and uses the L2 norm

$$E_{smooth\_Z} = \sum_{p\in\mathcal{P}} \left\| \nabla Z_f\left(\bar{x}_f\right) \right\|^2. \tag{2.17}$$

Equation (2.17) is an L2 norm, but there are a variety of commonly used regularization functions that have different characteristics, particularly in the region of discontinuities. The L1 norm is used later in this dissertation, and the results of using an L2 vs. an L1 norm are compared.

Combining (2.16) and (2.17) results in the energy functional

$$E_{total} = \gamma E_{match} + \bar{\alpha} E_{smooth} \tag{2.18}$$

where γ and $\bar{\alpha}$ are tuning constants and

$$\bar{\alpha} = \left(\alpha_x, \alpha_y\right)^T \tag{2.19}$$

allows tuning the smoothness along a radial epipolar line independently of the smoothness between radial epipolar lines.

Equation (2.13) represents two equations in two unknowns, one in the x direction and the second in the y direction with two unknowns Z and $\dot{Z}$. This equation pair contains Z in quadric form in the denominator of $\bar{c}\left(\bar{x}_f\right)$, which has implications for the numerical solution.

2.4 Numerical Solutions

Equation (2.18) defines a global energy that can be solved using a variety of techniques including variational methods [17], [25], [26], simulated annealing [27]-[29], cooperative methods [30], [31], and more recently graph cut techniques [31]-[35]. The use of variational methods was one of the first techniques used to solve global energy problems in early vision [26], [36], and they are still the most widely used technique for computing optical flow from image sequences [37]-[39]. One advantage to the variational approach is that the problem is specified in continuous (infinitesimal) terms.

The graph cuts method has been successfully used in many stereo correspondence-finding algorithms [33], [35], [40] that require the optimization of energy functionals where the objective of the optimization is to assign a label to each pixel, which results in a global cost minimum. For stereo correspondence finding, the labels are typically discrete disparity values. The energy-minimization problem is formulated as a 3D graph, with pixel locations in x and y and discrete disparities (labels) in z. The minimum cut (or maximum flow) of the 3D graph produces a dense disparity map, which is the global minimum of the energy functional. Section 2.6 of this dissertation applies this approach to (2.18) to find the global minimum, which results in the alignment of the optical flow fields.

Scharstein et al. [31] showed that for stereo correspondence finding, graph cuts produced the lowest error rates in terms of RMS disparity errors, number of bad pixels, and accuracy in textureless regions and was the second best global method in terms of accuracy in occluded areas. The graph cuts solution to the global energy-minimization problem of (2.18) is an NP-hard combinatorial problem, but highly efficient approximate

solutions [33] are widely used to find approximate solutions to energy functionals of the form of (2.18).

2.5 Variational Methods

Variational methods require that the energy be expressed in a continuous form such that the first variation can be found. Additionally, we want to separate Z from $\dot{Z}$ in the formulation to allow a gradient descent with respect to Z while lagging the solution for $\dot{Z}$.

2.5.1 Euler-Lagrange

We rewrite (2.16) and (2.17) in continuous form using the L2 norm for the regularization

$$E_{match} = \frac{1}{2}\int_a^b \left[m(\bar{x}_f)\bar{w}_f(\bar{x}_f) - c(\bar{x}_f)\bar{w}_b\left(\bar{x}_f + \bar{h}(\bar{x}_f)\right)\right]^2 d\bar{x} \tag{2.20}$$

$$E_{smooth\_Z} = \frac{1}{2}\int_a^b \left\|\nabla Z_f(\bar{x}_f)\right\|^2 d\bar{x} \tag{2.21}$$

where

$$m(\bar{x}_f) = \left(\frac{f_b}{f_f}\right)\left(\frac{Z(\bar{x}_f)}{(Z(\bar{x}_f)+b)}\right) \tag{2.22}$$

$$\bar{c}(\bar{x}_f) = \left(\frac{\bar{w}_f(\bar{x}_f)}{\left(\frac{Z_0(\bar{x}_f)+b}{Z_1(\bar{x}_f)+b}\right)\left(\frac{Z_1(\bar{x}_f)}{Z_0(\bar{x}_f)}\right)(\bar{w}_f(\bar{x}_f)+\bar{x}_f)-\bar{x}_f}\right). \tag{2.23}$$

We can now take the first variation of equations (2.20) and (2.21) with respect to Z

$$\gamma w_z(pw_l - w_r)(m^{'}w_f + mw_f^{'} - \bar{c}^{'}w_b\left(\bar{x}_f + \bar{h}(\bar{x}_f)\right) - \bar{c}w_b^{'}\bar{h}^{'}) - \nabla \cdot \begin{bmatrix} \alpha_x & 0 \\ 0 & \alpha_y \end{bmatrix} \nabla Z_1 \tag{2.24}$$

where

$$\bar{h}(\bar{x}_f) = \frac{\bar{x}_f\left(\frac{f_b}{f_f} - 1\right)}{Z+b} - \frac{\bar{x}_f\left(\frac{f_b}{f_f}Z - Z - b\right)}{(Z+b)^2} \tag{2.25}$$

$$m^{'} = \frac{\partial m}{\partial Z} = \left(\frac{f_b}{f_f}\right)\left(\frac{b}{(Z_1+b)^2}\right) \tag{2.26}$$

$$w_f^{'} = \frac{\partial w_f}{\partial Z} = -\frac{w_f}{Z_1} \tag{2.27}$$

$$w_b^{'} = \frac{\partial w_b}{\partial Z} = -\frac{w_b}{Z_1+b} \tag{2.28}$$

$$\bar{c}^{'} = \frac{\partial c}{\partial Z} = \left(\frac{\bar{w}_f\left({Z_0}^2 + bZ_0\right)\left(\bar{w}_f + \bar{x}_f\right)(2Z_1+b)}{\left[\left(\frac{{Z_0}^2+bZ_0}{{Z_1}^2+bZ_1}\right)\left(\bar{w}_f+\bar{x}_f\right) - \bar{x}_f\right]\left({Z_1}^2+bZ_1\right)^2}\right) \tag{2.29}$$

$$\nabla \cdot \begin{bmatrix} \alpha_x & 0 \\ 0 & \alpha_y \end{bmatrix} \nabla Z_1 = \alpha_x \frac{\partial^2 Z}{\partial x^2} + \alpha_y \frac{\partial^2 Z}{\partial y^2} \tag{2.30}$$

$$w_z = m(\bar{x}_f)w_f(\bar{x}_f) - \bar{c}(\bar{x}_f)w_b\left(\bar{x}_f m(\bar{x}_f)\right) \tag{2.31}$$

The solutions lie on radial epipolar lines (Figure 2.3). The Euler-Lagrange equations (one in the *x* direction and the other in the *y* direction) are solved using the gradient descent method.

(a) (b)

Figure 2.3. Radial epipolar line for coaxial camera rig. (a) Back camera. (b) Front camera.

2.5.2 Implementation Details

2.5.2.1 Discrete Laplacian

The discrete Laplacian is computed using a finite difference scheme.

2.5.2.2 Initialization

We initialize the value of Z by observing that the optical flow vectors that start and end on the optical axis (e.g., $\bar{X} + \Delta\bar{X} = 0$ or $\bar{X} = 0$) result in $c(\bar{x}_f) = 0$, allowing the computation of Z directly on the optical axis

$$Z(\bar{x}_f = (0,0)^T) = \frac{b}{\frac{w_f(0) f_b}{w_b(0) f_f} - 1}. \tag{2.32}$$

With $Z(\bar{x}_f = (0,0)^T)$, we can compute the velocity $\dot{X}$ and then use the projection equation and optical flow to estimate Z for all pixels in the images. For rigid scenes

with no Z translation, this method results in the initial estimate being a close approximation to the depth map if the optical flow fields are a good approximation of the projection of the motion field. Where there is a change in Z between consecutive images in the image sequence and/or where the scene is not rigid, this method produces a reasonable starting point for the gradient descent iterations.

2.5.2.3 Resampling to a Discrete Grid

The gradient descent results in a new estimate of Z at $t = n + 1$ after each step. This estimate, however, is offset spatially by the optical flow. Because optical flow algorithms produce subpixel flow values, the new Z values are rarely on integer pixel locations. This noninteger location requires resampling the newly estimated depth map onto an integer pixel grid to obtain the Z value that corresponds to each pixel. This linear interpolation resampling process introduces a slight smoothing to the depth estimation.

2.5.2.4 Stopping Criteria

We used one of two stopping criteria depending on the quality of the flow fields and the value chosen for $\bar{\alpha}$. When the flow fields closely represent the motion fields and $\bar{\alpha}$ is small (minimal Z smoothing), we compute

$$error_{flow\ match} = \left[m(\bar{x}_f)\bar{w}_f(\bar{x}_f) - \bar{c}(\bar{x}_f)\bar{w}_b\left(\bar{x}_f + \bar{h}(\bar{x}_f)\right)\right]^2 \qquad (2.33)$$

after each step in the gradient descent. Equation (2.33) is a measure of the mismatch in registration of the two flow fields. We stop iterating when (2.33) falls below a

predefined level; for the experiments with camera images, we used 0.01 pixels. Figure 2.4 shows the flow field alignment after each iteration of a typical gradient descent. For our experiments the gradient descent always stopped before 25 iterations.

Where the flow fields are noisy and not as good a representation of the motion field, larger $\bar{\alpha}$ values are typically required to get good results. With more substantial smoothing, the smoothing term (2.21) appears to pull the Z estimate away from the correct value if γ is large and/or if many iterations are performed. This effect is particularly evident around discontinuities in the scene. In this case we stopped the iterations when the smoothing term (2.21) was approximately equal to, but of opposite sign, the matching term (2.20). This latter approach produced larger residual values of w_z, but the experiments show that it results in more accurate depth estimations near discontinuities in the scene.

2.5.2.5 Algorithm

1) Compute $\overline{w}_f$ and $\overline{w}_b$.
2) Resample $\overline{w}_f$ and $\overline{w}_b$ along radial lines.
3) Smooth $\overline{w}_f$ and $\overline{w}_b$.
4) Initialize Z.
5) Iterate until stopping condition met.
    a) For each radial epipolar line:
        i) update Z estimate for one gradient descent step,
        ii) resample Z estimate to grid,
        iii) compute $\dot{z}$,

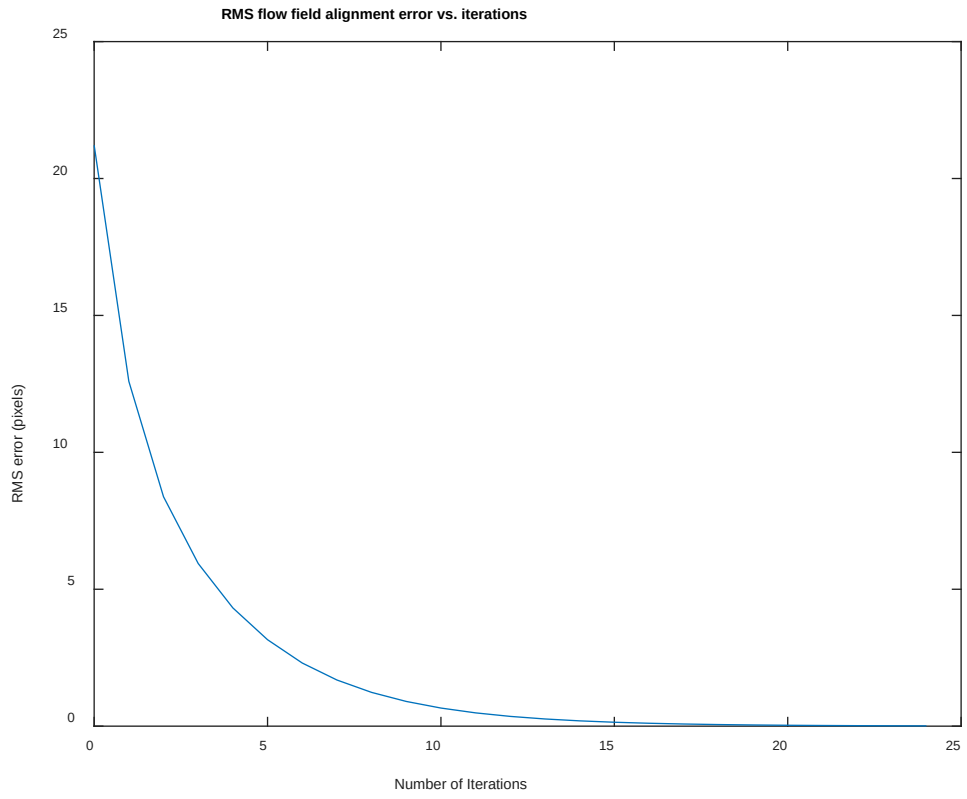


Figure 2.4. Typical RMS optical flow alignment error vs. gradient descent iterations.

iv) update $c(\bar{x}_f)$,

v) resample Z onto an XY grid.

2.5.3 Experimental Results

We tested the method on both synthetic optical flow fields as well as on real image sequences. The purpose of using the synthetic optical flow fields was to verify that the energy formulation, when solved, resulted in accurate depth estimations. If the optical flow fields are an accurate projection of the motion field and if the two cameras correctly perceive the motion field, then in all nonoccluded areas, the reconstructed depth map should be within a numerical estimation error of the ground truth.

2.5.3.1 Synthetic Optical Flow Fields

For the synthetic optical flow fields, we defined the geometry of a 3D scene and projected the 3D motion of that scene onto a virtual image plane via an ideal pinhole camera model (Figures 2.5 and 2.6), which results in a simulated optical flow field that is the projection of the 3D motion field. Additionally, the differences in the flow fields perceived by the two cameras are related by the projection equations used to derive the energy that we are minimizing. There is a discretization effect of converting the continuous flow field into discrete pixel locations, which results in small interpolation errors when resampling onto a pixel grid after each step of the gradient descent. Thus, the simulated flow field experiments provide an estimate of the upper boundary of numerical estimation precision for the methodology. We determine the accuracy of the resulting image alignment by estimating the depth map along radial epipolar lines and compare that to the original scene geometry by computing the RMS depth and disparity error.

For the synthetic flow images f_f = 4.8 mm, f_b = 4.0 mm, the cameras have .002 mm square pixels, velocity in the XY plane was varied from 0.5 m/s to 3.5 m/s, and velocity along the Z-axis ranged from 2.5 m/s toward the camera to 2.5 m/s away from the camera. The camera frame rate was set to 30fps. We set $\gamma = 1 \cdot 10^{11}$ and $\bar{\alpha} = [5 \cdot 10^{-5},\ 5 \cdot 10^{-5}]$.

Figures 2.7(a) and 2.8(a) show the results for a smooth scene for a horizontal line. With the exception of the slowest XY displacement (0.5 m/s) and highest Z displacements, the RMS depth error is < 0.15%. The shape of the curves suggests that larger displacements in the Z direction produce less accurate results, likely due to lagging

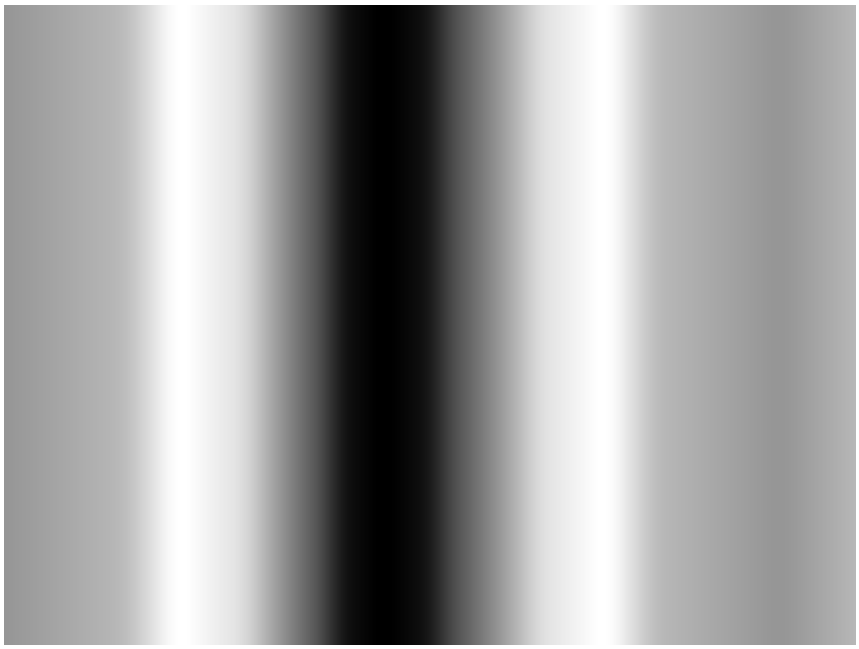


Figure 2.5. Depth map for synthetic image of smooth scene.

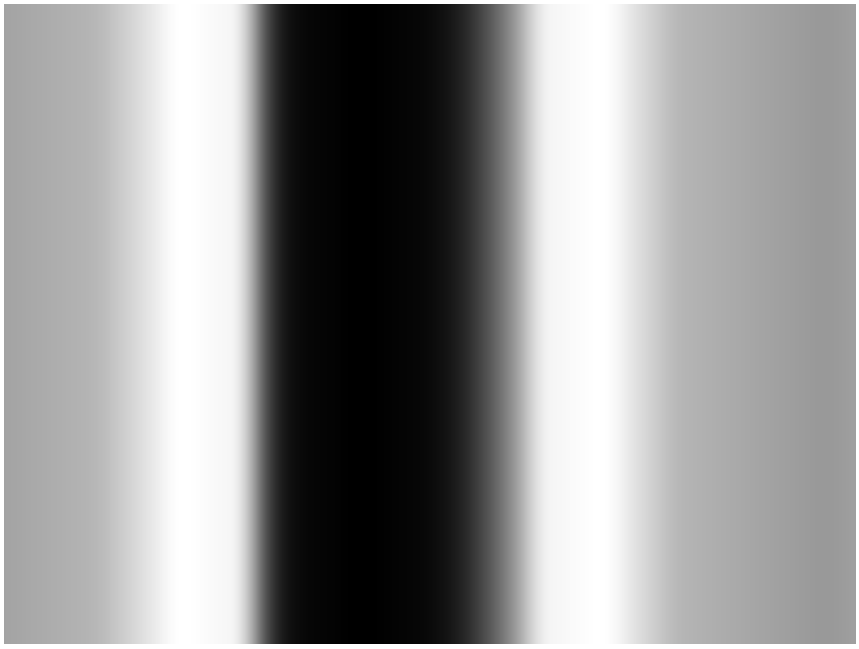
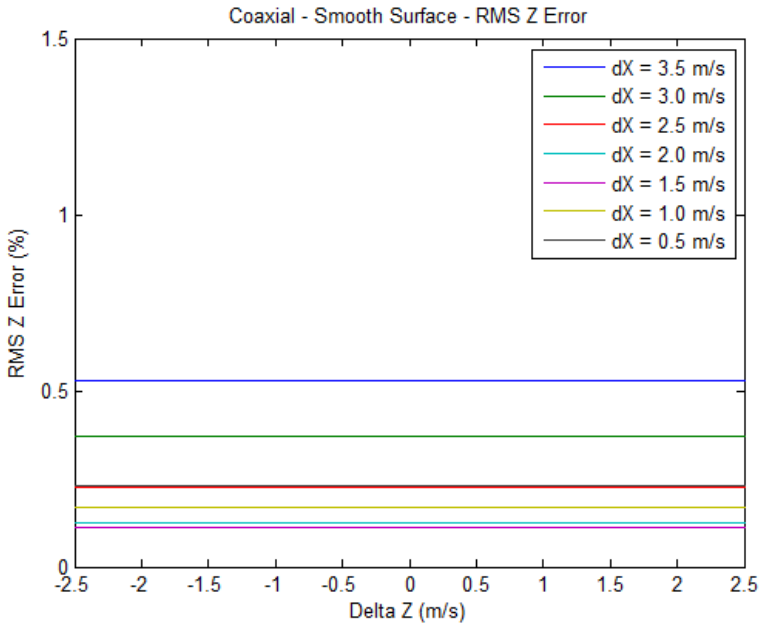
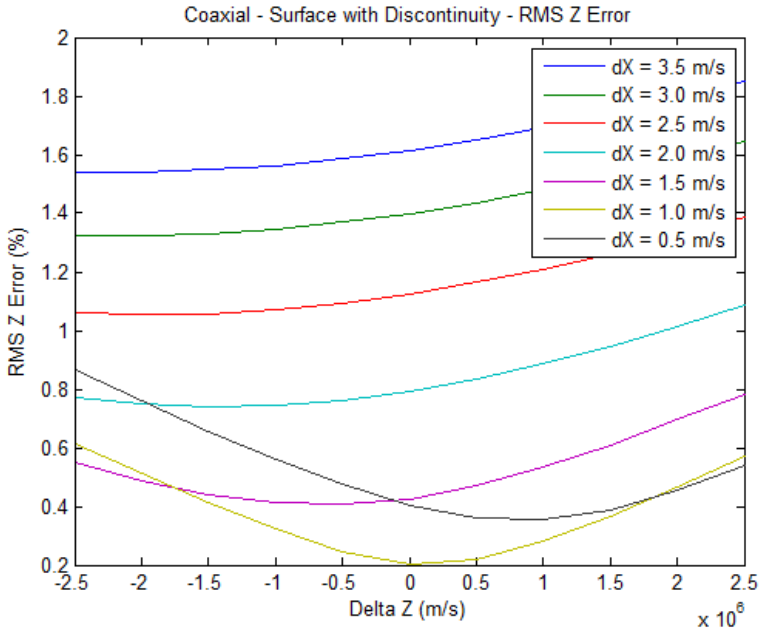


Figure 2.6. Depth map for synthetic image of scene with occlusions.

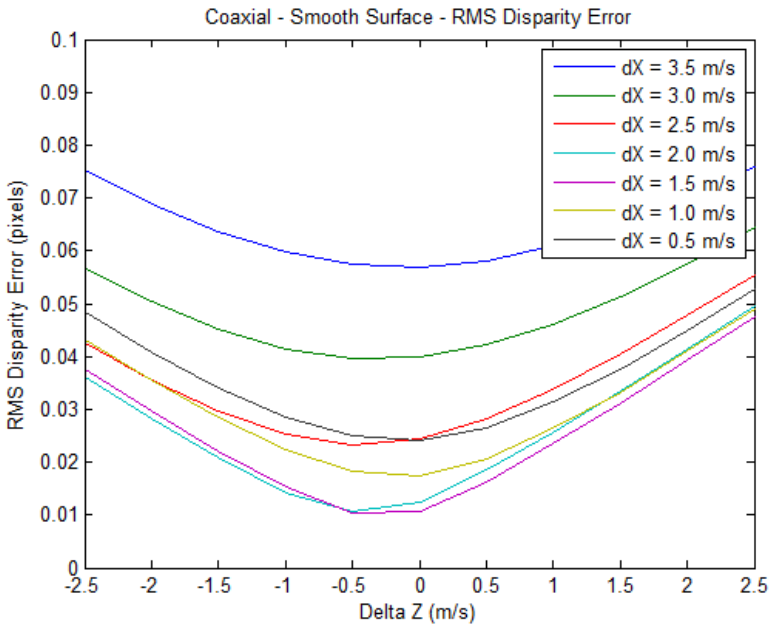


(a)

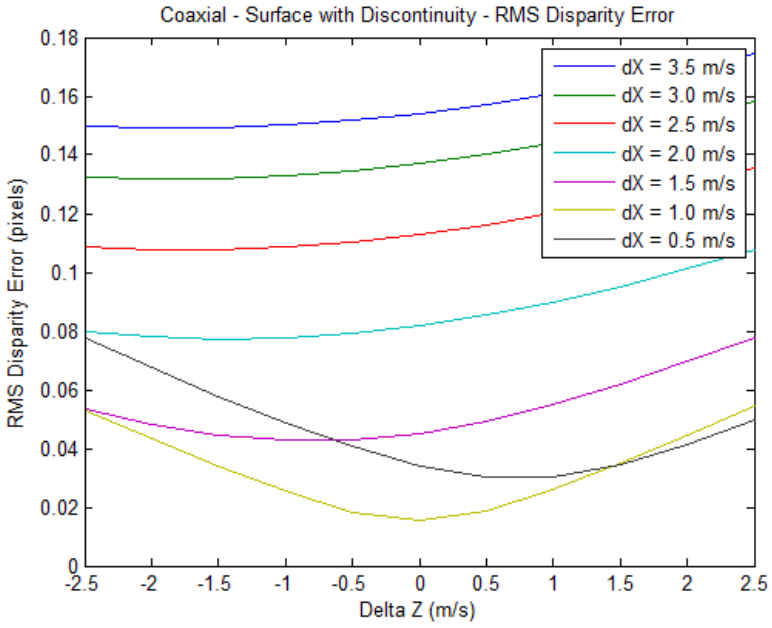


(b)

Figure 2.7. RMS Z error for coaxial camera rig using synthetic flow fields. (a) Smooth surface. (b) Surface with discontinuities and occlusions.



(a)



(b)

Figure 2.8. RMS disparity error for coaxial camera rig using synthetic flow fields. (a) Smooth surface. (b) Surface with discontinuities and occlusions.

the solution for $\dot{Z}$, which we later resolved by simultaneously solving for Z and $\dot{Z}$ in the graph cuts approach (see Section 2.6). $\dot{Z}$ produces flow along radial lines whereas $\dot{X}$ and $\dot{Y}$ produce horizontal and vertical flow. Where flow due to $\dot{X}$ and $\dot{Y}$ is aligned with radial lines, lagging the solution for $\dot{Z}$ combines the flow due to $\dot{Z}$ with that due to $\dot{X}$ and $\dot{Y}$. This issue can be resolved by rotating the flow field into one component along each radial line and the other component perpendicular to the radial line. This method isolates the flow due to $\dot{X}$ and $\dot{Y}$ from the flow due to $\dot{Z}$; however, the more computationally efficient solution is to solve simultaneously for Z and $\dot{Z}$.

Figures 2.7 (b) and 2.8 (b) show the results for a synthetic flow scene with a large discontinuity that produced occluded areas. As expected, the RMS errors increase, but the increase is modest, and one would expect it to be smaller than the RMS errors from comparable binocular stereo camera rig due to the smaller occlusions. This result is confirmed in the experiments with real images when comparing the coaxial camera rig results to those of a multimodal stereo camera rig.

2.5.3.2 Flow Fields From Camera Images

The coaxial camera rig (Figure 2.9) consists of a pair of Point Gray 0.3MP Color Firefly MV 1/3" CMOS computer vision cameras with global shutters. The camera rig was mounted on an 8" x 8" optical breadboard with micrometer adjustable rotational stages to allow precise alignment of the optical centers. We used a 50/50 plate beam splitter from Edmund Optics part number 46-583. In Figure 2.9 the light baffle on the beam splitter has been removed for clarity.

The camera rig was mounted on a precision ball lead screw XY table with 800 count

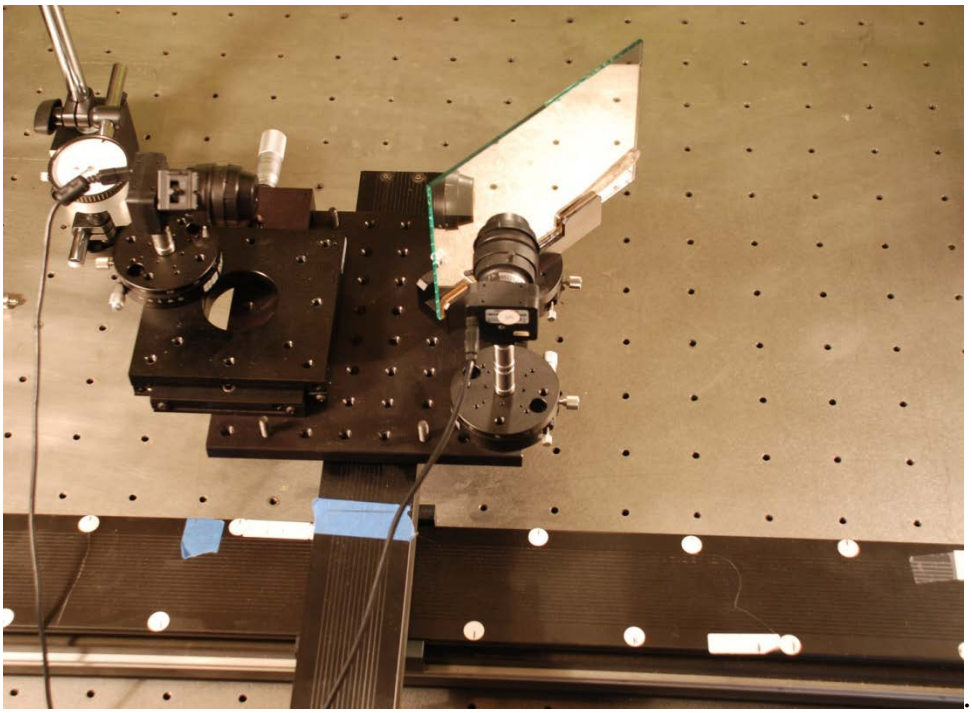


Figure 2.9. Coaxial camera rig on XY table.

per rotation optical encoders. The X axis of the table has a linear resolution of 0.025mm, and the Y axis of the table (which is the Z axis in images) has a resolution of 0.0125mm. The XY table allowed the camera rig to be translated a known distance between frames.

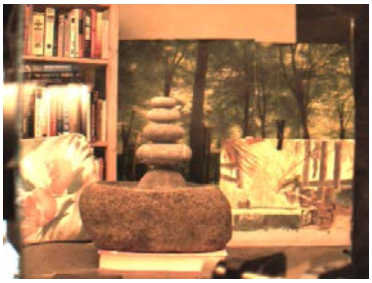
The cameras were calibrated using Cal Tech's Camera Calibration Toolbox [41] based on the work of Zhang et al. [42], [43]. Both cameras were calibrated through the beam splitter. The image received by the front camera was mirrored to have the same orientation as the back image (which is mirrored due to the beam splitter).

The scenes are shown in Figures 2.10, 2.11, and 2.12 (a), (b), (d), and (e), and the resulting optical flow in (c) and (f). The first scene (Figure 2.10) consists of a stone fountain located about 90 cm from the optical center of the front camera in the camera rig and several background objects located between approximately 120 cm and 250 cm from

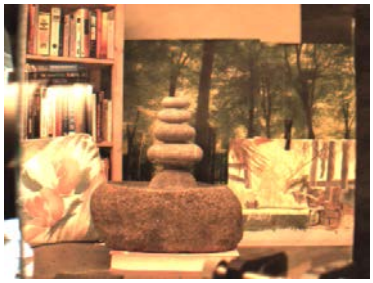
(a)

(b)

(c)



(d)

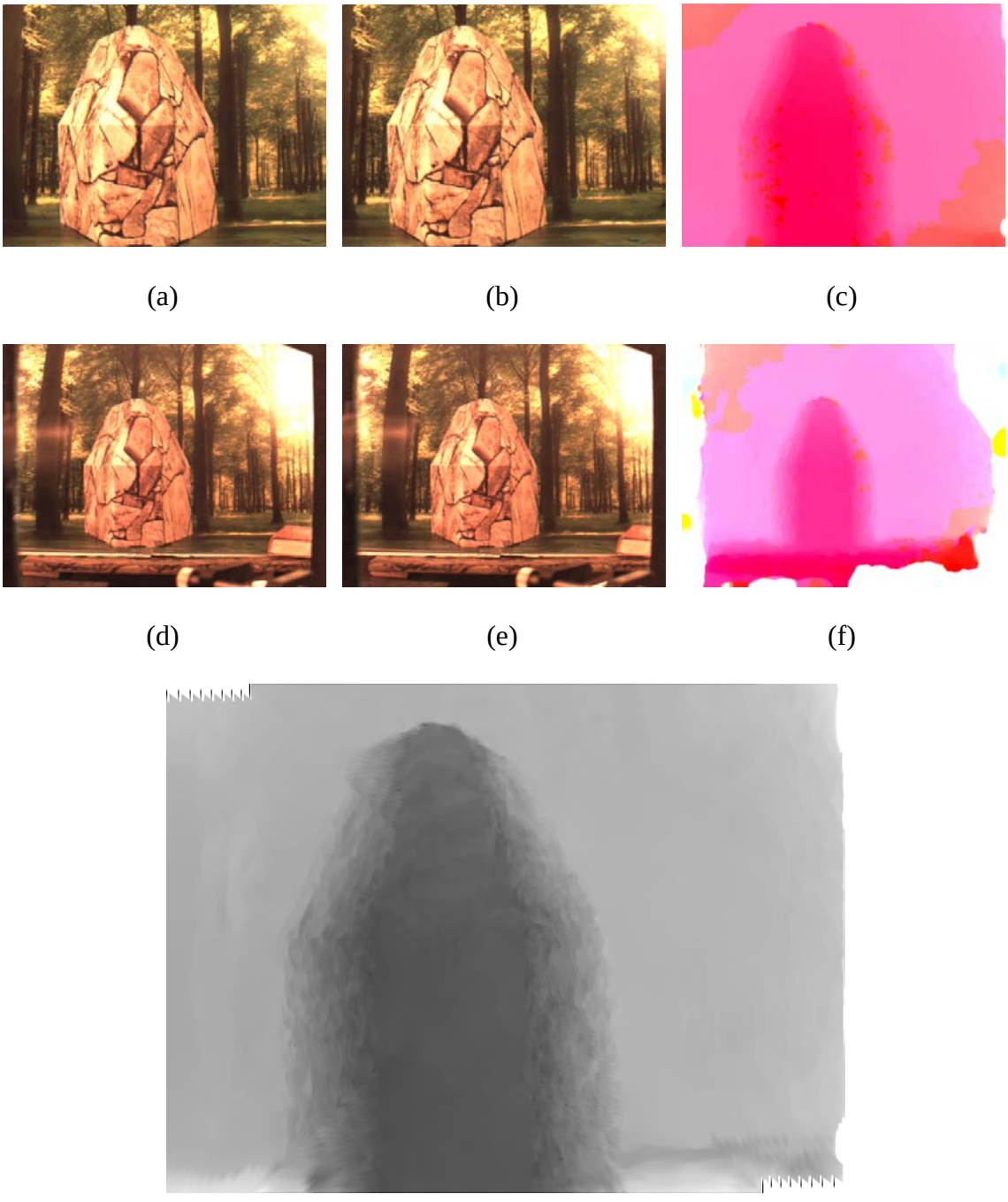


(e)

(f)

(g)

Figure 2.10. Fountain image sequence, coaxial camera rig, variational methods: (a) first front camera image, (b) second front camera image, (c) optical flow from front camera, (d) first back camera image, (e) second back camera image, (f) optical flow from back camera, (g) resulting depth map using variational methods.



(a) (b) (c)

(d) (e) (f)

(g)

Figure 2.11. Flagstone image sequence, coaxial camera rig, variational methods: (a) first front camera image, (b) second front camera image, (c) optical flow from front camera image pair, (d) first back camera image, (e) second back camera image, (f) optical flow from back camera image pair, (g) resulting depth map.

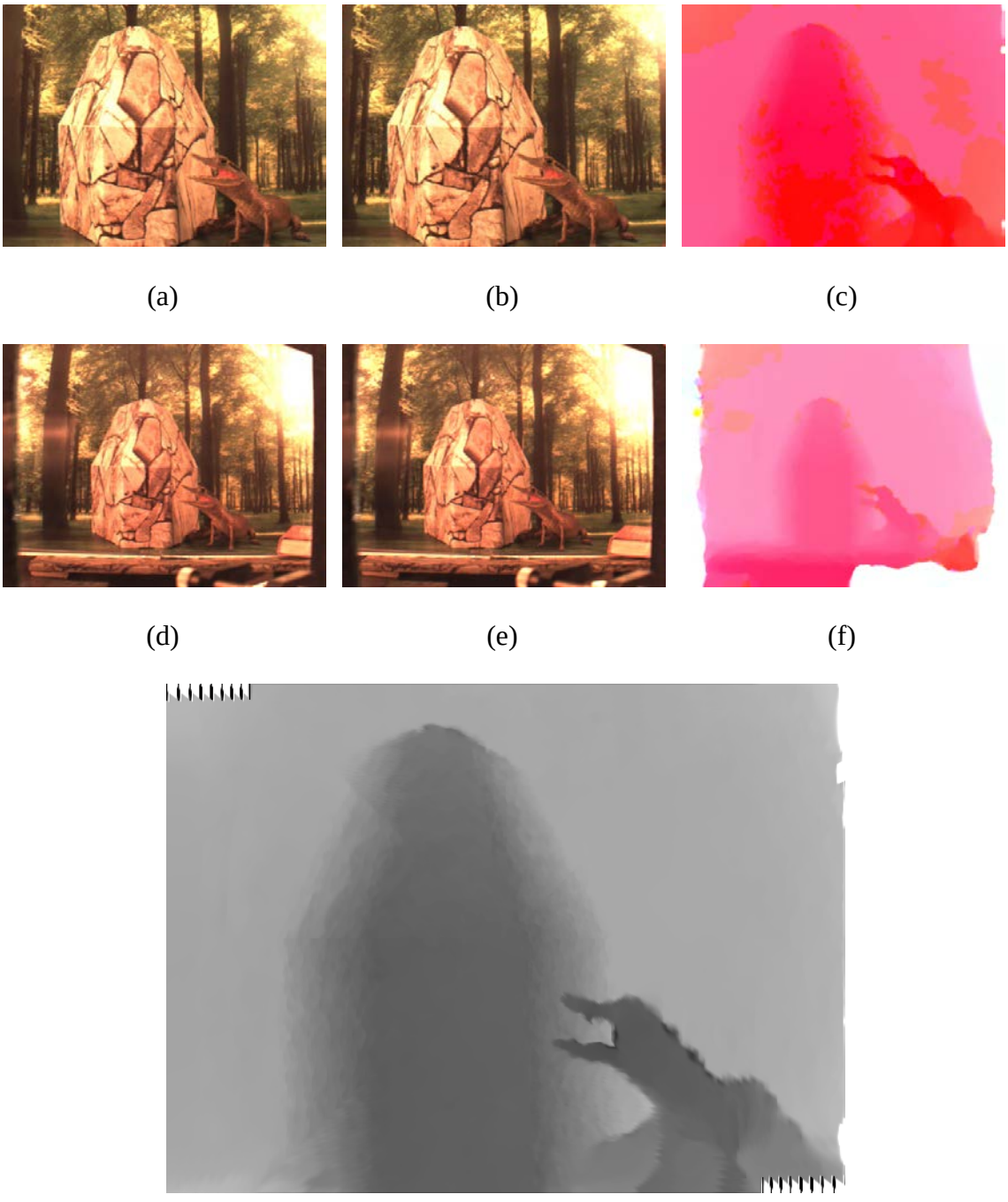


Figure 2.12. Flagstone with alligator image sequence, coaxial camera rig, variational methods: (a) first front camera image, (b) second front camera image, (c) optical flow from front camera image pair, (d) first back camera image, (e) second back camera image, (f) optical flow from back camera image pair, (g) resulting depth map.

the optical center of the front camera. The purpose of this scene was to see how the methodology performed on a scene with significant occlusions.

The second scene (Figure 2.11) consists of a geometric shape against a frontal planar background. This scene was specifically constructed to avoid occlusions and to contain large frontal planar surfaces.

The third image scene (Figure 2.12) added a small alligator into the previous scene. The alligator was intended to explore how the methodology handles fine features, but it also adds some smaller occlusions.

The camera rig was translated 20 mm between image frames, which equates to a velocity of 0.6 m/s for a 30 fps frame rate. The cameras have 0.006 mm square pixels, focal lengths of 7.7 mm and 5.8 mm (front and back, respectively), and the baseline b = 143.3 mm. We set $\gamma = 2 \cdot 10^8$ and $\alpha = [.01, .001]$. We used the large-scale optical flow algorithm from [24]. The flow in the x and y directions was resampled to radial epipolar lines at one degree increments, which provides dense reconstruction near the center of the image, but leaves some small gaps near the edges of the images, which we approximated by interpolation between epipolar lines when converting the depth along epipolar lines back to an XY grid. Some small shark-tooth-like anomalies due to the horizontal line to radial line resampling can be seen in the dense depth map. These are visible in the upper left and lower right corners of the reconstructed depth map.

We compare the results using our method -- image correspondences from perceived motion (ICPM) -- with those from a state-of-the-art two-view scaled shape from motion (SfM) algorithm. SfM finds a set of set of common features between image pairs. We detect corners using a minimum eigenvalue algorithm, and then we use the Kanade-

Lucas-Tomasi (KLT) tracker to track the movement of these features. From the feature matches, we can estimate the fundamental matrix and compute the relative camera poses. This method allows us to compute the depth of the matched points using triangulation, which gives us a sparse depth map up to scale. The scale is recovered using a known scene dimension.

Table 2.1 shows the flow field alignment errors in pixels for each scene along with the RMS error of the reconstructed camera movement using both ICPM and SfM. The reconstructed camera movement was computed from the estimated Z and $\dot{Z}$ and the computed optical flow w_x and w_y for each nonoccluded pixel using (2.6) and (2.7). ICPM produces 2.5% to 73% reduction in the error of the reconstructed camera motion vs. SfM as well as fully dense depth maps vs. the sparse depth maps that result from using scaled SfM. Additionally, ICPM does not require knowledge of a scene dimension to obtain scale, but it does require two cameras.

Figures 2.10(f), 2.11(f), and 2.12(f) show the dense depth maps using ICPM. In the dense depth maps, the closer the object is to the camera, the darker the pixel.

Several categories of visual anomalies can be seen in the dense depth maps. The

Table 2.1. Coaxial variational methods: alignment errors, scene flow errors, and computational time.

	Fountain	Flagstone	Flagstone + alligator
ICPM Flow Alignment RMS Error	<0.01 pixels	<0.01 pixels	<0.01 pixels
ICPM Scene flow error	3.9%	3.9%	3.1%
Scaled SfM Scene flow error	14.4%	4.0%	3.6%
ICPM Computational Time	21.1 seconds	19.9 seconds	21.0 seconds
SfM Computational Time	4.4 seconds	10.1 seconds	5.1 seconds

most visible anomaly is the mottled appearance of planar surfaces. This anomaly comes from the estimate of the optical flow, which has the same mottled appearance. This visual anomaly can be eliminated by increasing the weight of the smoothing terms, but at the expense of losing finer details.

The second category of anomalies includes those due to resampling between the radial lines and the XY grid. The most visible are the small shark-tooth-like features in the upper left and lower right corners of Figure 2.10(f) and Figure 2.12(f). These features are caused by resampling at one-degree increments, creating a 5-pixel gap between the radial lines at the edge of the image.

The third category of anomalies is related to occlusions. The occluded areas for the coaxial camera rig are very small, but they are still visible in Figure 2.10(f) around the edges of the fountain and along the left edge of the frontal planar surface behind the fountain to the right. Sometimes the algorithm estimates a good value for the occlusions, but in other areas the depth estimate in the occluded area either shows up too low (e.g., darker than the surrounding area) or too high (lighter than the surrounding area). As we will see in the next section, these occluded areas are substantially smaller for the coaxial camera rig than for a similar geometry multimodal stereo camera rig, but they still exist.

2.5.4 Discussion

From the simulated flow experiments we learn two things. First, the methodology is capable of aligning flow fields in such a way that the underlying images are also aligned. Second, the aligned flow fields can be used to estimate dense depth maps without using intercamera image correspondences. This method works equally as well in the center

region of a coaxial camera where disparity is zero or close to zero, as it does in the outer regions of the images. This result demonstrates that it is possible to resolve the unrecoverable point problem first described by Ma and Olsen [15] 25 years ago.

Errors in the simulated flow field experiments are below those of real-world optical flow algorithms, which suggests that the limitation of this methodology, in terms of accuracy, will be in the optical flow computation, not in the flow field alignment.

For the real-world images, we see equally good alignment of the flow fields, but with some visual anomalies in the estimated depth maps due to the limited resolution and smoothing of the optical flow fields.

Depending on the scene, our method is slightly more accurate (2.5% reduction in scene flow error) to substantially more accurate (73% reduction in scene flow error) when compare with scaled SfM. In addition, scaled SfM is a sparse technique finding matches for less than 2% of the pixels in the test images and requiring at least one known dimension in the scene in order to scale the depth maps.

The results of both the simulated optical flow field experiments and experiments performed on real images are promising, but several limitations are associated with solving the energy functional using variational methods. First, the number of iterations required to come to a good solution is dependent on the quality of the initial estimate. The computational time is directly proportional to the number of iterations. We were able to make good initial estimates in the experiments, which kept the number of iterations to 25 or fewer, but for poor initial estimates, for which hundreds of iterations may be required.

Second, determining a good stopping condition is problematic. In image sequences

with occlusions, there is contention between the matching term and the smoothness term, particularly in the areas of occlusions. Increasing the number of iterations causes the error term (2.31) to approach zero; however, after a certain number of iterations, the resulting depth map becomes less accurate in the region around occlusions. This result appears to be due, at least in part, to the interpolation that occurs after each iteration to realign the depth map with the pixel grid. The worst scene flow error was for the fountain image, which has large occlusions. The source of the scene flow error is primarily due to how the optical flow algorithm handles the large occlusions, and thus one would expect similar errors in decoupled scene flow algorithms when there are large occlusions.

Third, lagging the solution for $\dot{Z}$ appears to be a source of error in finding the disparity and depth, likely due to the effect that motion in the Z direction produces flow in the x and y directions. Mathematically, the effect of $\dot{Z}$ can be isolated by resampling the x and y components of the optical flow in one component of the optical flow along each radial line and the second component of the optical flow perpendicular to each radial line. The component of the optical flow perpendicular to radial lines is due only to scene flow in the X and Y directions. However, this addition further increases the computational complexity. The better approach is to solve simultaneously for Z and $\dot{Z}$, which as we will see, is done more effectively using an optimization technique based on graph cuts.

Lastly, the initialization procedure can be problematic when $\dot{Z}$ is large relative to $\dot{X}$ and $\dot{Y}$. As in the case of the problems caused by lagging the solution for $\dot{Z}$, initialization could be improved by rotating the flow field components into radial flow and flow

perpendicular to the radial line, but as described above, this increases the computation complexity and is better solved using an optimization technique that does not require initialization.

Even with the above-mentioned problems, the reconstructed 3D depth maps are visually realistic and the accuracy is better than the current state of the art, which suggests that the coaxial camera can be a valuable tool in computer vision applications where a binocular stereo rig does not work well. As we will see in the next section, using graph cuts to solve the energy-minimization problem has little effect on the quality and accuracy of the resulting depth map; however, graph-cuts-based optimization does not require initialization (other than selecting a finite list of labels), solves for Z and $\dot{Z}$ simultaneously, and is less sensitive to the value of the tuning constants.

2.6 Graph Cuts

Graph cuts have been effectively used to solve a number of energy-minimization problems related to early vision that can be written in the form

$$E(\mathcal{L}) = \sum_{p \in \mathcal{P}} D(\mathcal{L})_{data} + \sum_{p \in \mathcal{P}} V(\mathcal{L})_{smooth} \tag{2.34}$$

where $\mathcal{L}$ is a finite set of labels, $D(\mathcal{L})_{data}$ is a data matching energy term, $V(\mathcal{L})_{smooth}$ is a smoothness term, and $E(\mathcal{L})$ is the total global energy to be minimized. In this section we will give a brief background on using graphs to solve min-cut/max-flow problems in early vision. Next we describe the Boykov-Kolmogorov algorithm that we use to solve the energy minimization. Lastly, we discuss several important implementation details

including the computation of data costs, the construction of the labels matrix, and the neighborhood structure.

2.6.1 Background on Graphs

In network flow problems, graph theory is the study of graphs, which consist of a set of nodes or vertices, $\mathcal{V}$, connected by arcs or edges, $\mathcal{E}$. The graph is an ordered pair of vertices, and edges, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Each edge is an ordered pair of two vertices (p, q). Ordered pairs of vertices are assigned edge costs or edge weights. If the cost between vertices (p, q) is the same as the cost between (q, p), then the graph is called undirected. If the cost depends on the order of the vertices, then the graph is called directed.

Graphs typically contain two special vertices (terminals) called the sink, t, and the source, s. In computer vision problems the vertices are typically pixels and the edges represent the pixel neighborhood.

2.6.2 Min-cut and Max-flow Problems

In graph theory, a cut partitions the vertices into two subsets, $\mathcal{S}$ and $\mathcal{T}$, where $\mathcal{S}$ contains the source terminal s and $\mathcal{T}$ contains the sink terminal t. This partition is called an *s/t* cut, $C = \{\mathcal{S}, \mathcal{T}\}$. The cost of a cut, C, is the sum of the costs of all the edges that link a vertex in $\mathcal{S}$ to a vertex in $\mathcal{T}$. A minimum cut is the partition of vertices into two disjoint sets that produce the minimum cost.

A min-cut problem can also be formulated as a max-flow problem where each edge has a maximum flow capacity that can pass through the edge. With the exception of the source and sink terminals, each vertex must have the same flow into and out of the

vertex. This restriction is called the conservation of flow constraint. The source terminal only has flow out and the sink terminal only has flow in. The Max-flow, Min-cut theorem of Ford and Fulkerson [44] states that the maximum flow from s to t saturates a set of edges. This set of saturated edges partitions the vertices in two disjoint sets, $\mathcal{S}$ and $\mathcal{T}$, which is the same partition that produces the minimum cut.

Min-cuts (or max-flows) can be applied to solve a number of early vision energy-minimization problems. A minimum cut partitions a group of pixels (vertices) into two disjoint sets: one containing the source and one containing the sink along some minimum global energy. For stereo correspondence finding, the graph can be thought of as a 3D cube with the x and y dimensions being the pixels in the image and the z dimension being disparity; thus each vertex represents a pixel at a specific disparity. An *s/t* cut is then a surface that partitions the pixels/disparity combination along a disparity surface, which produces the minimum global energy.

2.6.3 Boykov-Kolmogorov Algorithm

Numerical solutions to min-cut/max-flow problems fall into one of two main groups: augmenting path methods and preflow-push (or push-relabel) methods. Augmenting path algorithms, based on the original Ford-Fulkerson approach, perform a global augmentation by pushing flow into paths between the source and sink that are not yet saturated. In push-relabel algorithms the flow is pushed along individual edges. This step violates the conservation of flow constraint during intermediate stages of the algorithm, but generally produces a more computationally efficient result.

The Boykov-Kolmogorov algorithm [33], [40], [45], [46] is based on the augmenting

path algorithm, but with three main differences. Unlike traditional augmenting path algorithms, which build a breadth-first search tree from the source to the sink, the Boykov-Kolmogorov algorithm builds two search trees, one from the source to the sink and a second from the sink to the source. The second difference is that the Boykov-Kolmogorov algorithm reuses the search trees instead of rebuilding them after each path of a certain length is saturated. Rebuilding the search trees is a computationally expensive component of the algorithm as it involves scanning the majority of pixels in the images. The third difference is that the Boykov-Kolmogorov algorithm uses one of two different moves depending on whether the smoothing term is a metric or semimetric. If the smoothing term is a semimetric, then an α-β swap move is used, whereas if the smoothing term is a metric, then an α expansion move is used. These two move types allow simultaneously changing labels for a large set of pixels. According to Boykov and Kolmogorov [33], the α expansion algorithm finds a labeling within a known factor of the global minimum. Because our smoothing term is a metric, we use the α expansion algorithm.

In the Boykov-Kolmogorov algorithm, the two search trees consist of active and passive vertices. Active vertices are those that can grow, but passive vertices cannot grow because they are blocked by the surrounding nodes. The algorithm iterates through three stages, the grow stage, the augmentation stage, and the adoption stage.

In the grow stage of the algorithm, paths are grown from both the source and the sink. Growth occurs into all neighboring active vertices using nonsaturated edges. This stage stops when an active vertex from one tree encounters a neighboring vertex in the other tree. The result of the grow stage is a path from the source to the sink.

In the augmentation stage, we push through the largest flow possible along the path between the source and the sink. This stage generates a certain number of saturated edges. Saturated edges typically result in some vertices becoming "orphans." An orphan has been disconnected from the trees that start from the source and the sink terminals and becomes the root of a new tree. These new trees, however, do not contribute to the flow between the source and sink.

In the adoption stage the two-tree structure (one with the source as its root and one with the sink as its root) is restored. This restoration is done either by finding a valid parent for the orphans, or if a valid parent cannot be found, by removing the orphans.

The algorithm repeatedly iterates through the three stages until the two trees can no longer grow, and all the edges that connect the two trees are saturated. The fact that all the edges that connect the two trees are saturated implies that this is a maximum flow. In tests performed by Boykov and Kolmogorov, their algorithm performed two to five times faster than other methods.

2.6.4 Implementation Details

2.6.4.1 Veksler-Delong Implementation

This work relies on a publicly available version of the Boykov-Kolmogorov algorithm implemented by Veksler and Delong [47]. This version has a MATLAB wrapper, which may explain the slightly slower than expected computational performance. Using Veksler and Delong's notation, and referring to (2.34), $\mathcal{P}$ is a set of observations (e.g., pixels) and $\mathcal{L}$ is a finite set of labels (e.g., disparity values in traditional binocular stereo correspondence finding). D computes the cost of assigning a

particular label ℓ to pixel $\mathscr{p}$, and V is a regularization term that favors spatial smoothness. The objective is to assign each observation p a label ℓ such that the sum over all pixels $\mathcal{P}$ minimizes the global energy $E(\mathcal{L})$.

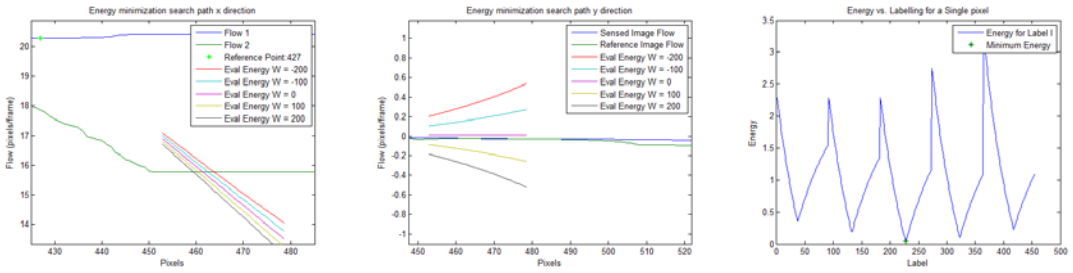
$\mathcal{L}$ is the finite set of $(Z, \dot{Z})$ pairs defined as

$$\mathcal{L} = \{(Z_{min}, -\dot{Z}_{min}), (Z_{min}+1, -\dot{Z}_{min}), (Z_{min}+2, -\dot{Z}_{min}), \dots, (\dot{Z}_{min}, -\dot{Z}_{min}+1), (Z_{min}+1, -\dot{Z}_{min}+1), (Z_{min}+2, -\dot{Z}_{min}+1), \dots, (Z_{max}, -\dot{Z}_{max}+1)\} \quad (2.35)$$

The matching term in this work (2.16) penalizes the difference between the optical flow in the reference image at pixel $\mathscr{p}$ and the optical flow in the sensed image at $\mathscr{p} + \bar{h}(\mathcal{L})$ when the optical flow is adjusted for the difference in magnification, which depends on the ratio of the focal lengths in the two systems and the ratio of the different Z distances of the two cameras. This is a two-component penalty as both components (w_x, w_y) of the optical flow contribute to the cost.

$E(\mathcal{L})_{smooth}$ is the sum over all pairs of neighboring pixels $(\mathscr{p}, \mathscr{q})$ in the reference image, where $(\mathscr{p}, \mathscr{q})$ are 4-connected. This cost defines the pixel neighborhood structure and assigns a linear penalty (L1) to neighboring pixels that have different labels.

The global energy has two notable differences when compared to a traditional binocular stereo energy: 1) in matching optical flow, each pixel location in the reference frame has two values, one for the optical flow in the x direction and a second for the optical flow in the y direction, and 2) Z and $\dot{Z}$ are solved for directly and simultaneously. This methodology can be visualized by computing the energy for a single point in the reference image (point 427 in Figure 2.13(a)) for a set of labels and observing the



(a) (b) (c)

Figure 2.13. Example energy response for a single pixel location.

response of the components of the energy. Figure 2.13(a) shows the pixel under evaluation in the sampled image (the green '*' on the blue curve labeled 'Flow 1'), the flow in the reference image (the green line labeled 'Flow 2'), and a series of straight lines that cross the flow from the reference image. Each of the straight lines represents a different value of $\dot{Z}$, and each point along a given straight line is a different value of Z. Notice that there is one minimum for each value of $\dot{Z}$. Figure 2.13(b) shows the same evaluation for the flow in the y direction and Figure 2.13(c) shows the combined energy for each label with a single minimum value. In this way we are able to solve for two unknowns (Z and $\dot{Z}$) with two equations (one for flow in the *x* and one for flow in the *y*).

2.6.4.2 Algorithm

1) Compute $\overline{w}_f$ and $\overline{w}_b$.
2) Resample the optical flow fields along radial epipolar lines.
3) Construct the cost matrix.
4) Construct the neighborhood matrix.
5) Define the smoothness costs.

6) Find the minimum cut.
7) Resample the optical labeling matrix back onto an XY grid.

2.6.5 Experimental Results

We tested the method on images from the same three scenes that we used with the variational methods solution. The scenes are again shown in Figures 2.14, 2.15, and 2.16 (a), (b), (d), and (e) and the resulting optical flow in (c) and (f). The reconstructed depth maps are in (g). Table 2.2 shows the alignment errors, scene flow errors, and computational time. Visually, the depth maps are equally realistic, but with some blockishness, a characteristic that is common when using graph cuts with L1 regularization. As with the depth maps from the variational methods approach, there are also some anomalies due to the conversion between radial lines and an XY grid. Alignment errors are similar to those found using variational methods.

2.6.6 Discussion

The visual results of using graph cuts to solve the energy-minimization problem are similar to the variational methods approach; however, the L1 norm that was used with graph cuts combined with the discrete labels results in a somewhat blockish visual appearance of the depth maps. Graph cuts has both advantages and disadvantages that make it more or less appropriate depending on the application and the nature of the scene. First, the graph cuts algorithm is slower, which is somewhat unexpected. This result is likely due to the variational methods approach using an initial estimate, which significantly reduces the number of iterations required to come to a "good-enough"

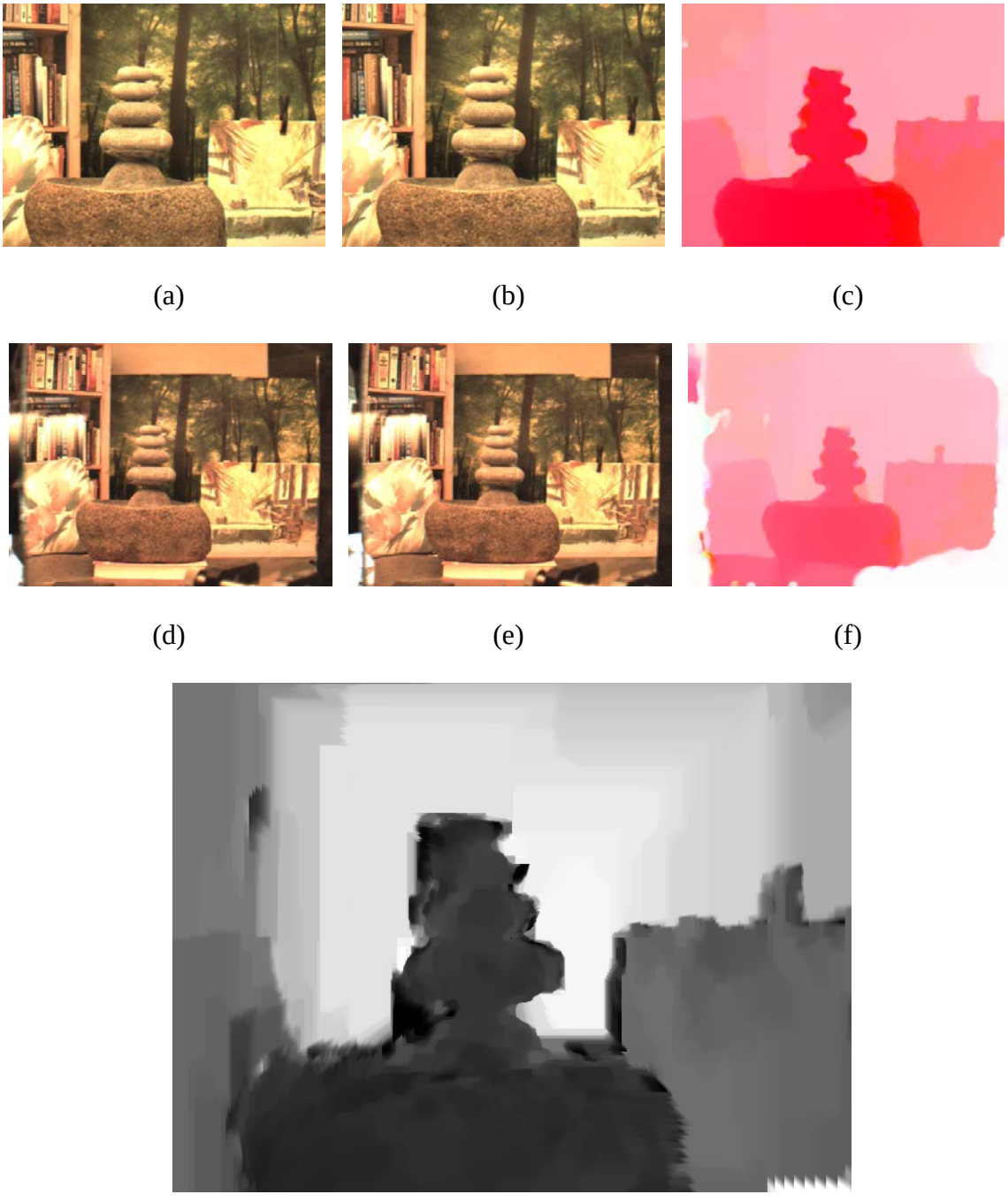
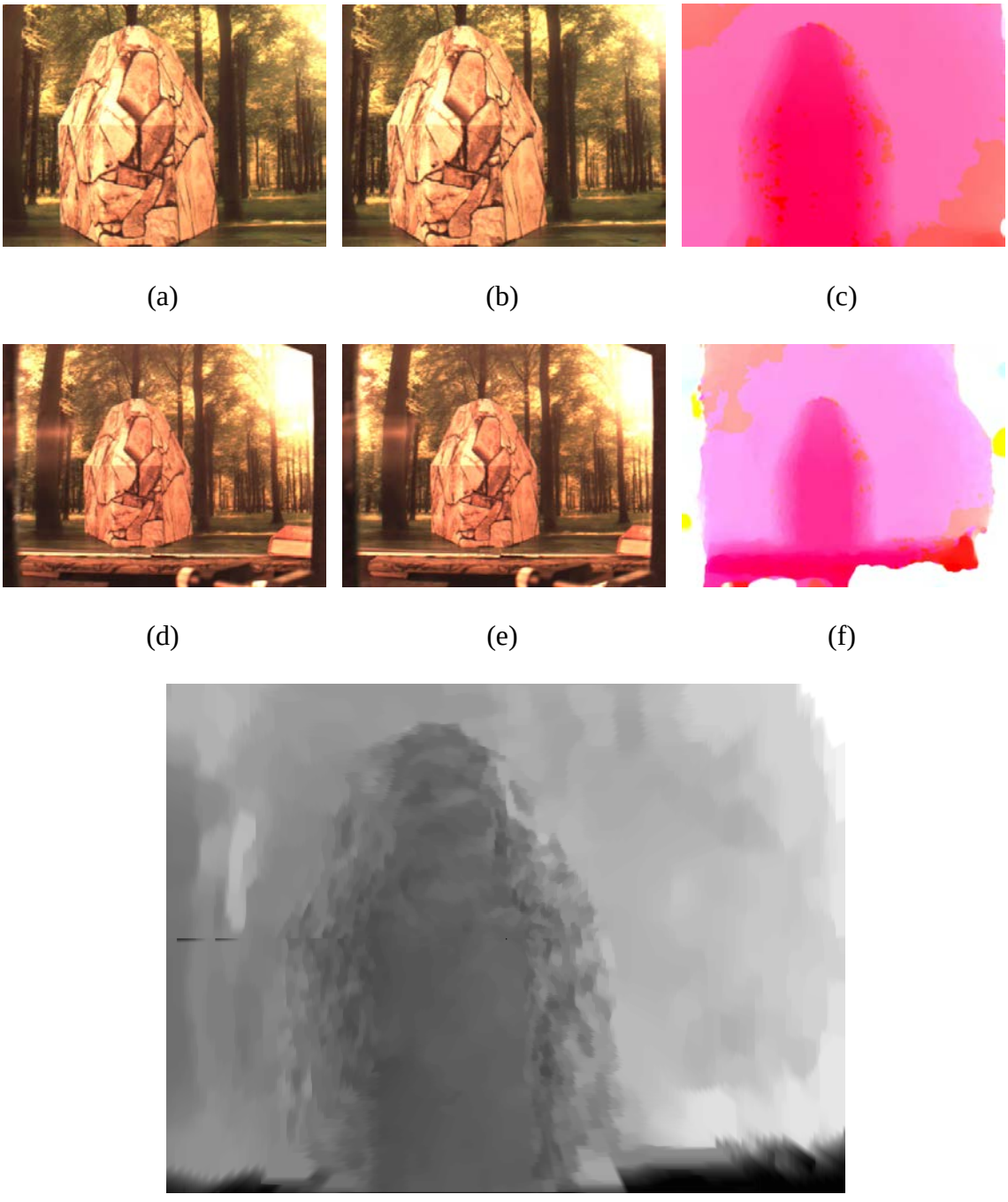


Figure 2.14. Fountain image sequence, coaxial camera rig, graph cuts: (a) first front camera image, (b) second front camera image, (c) optical flow from front camera image pair, (d) first back camera image, (e) second back camera image, (f) optical flow from back camera image pair, (g) resulting depth map.

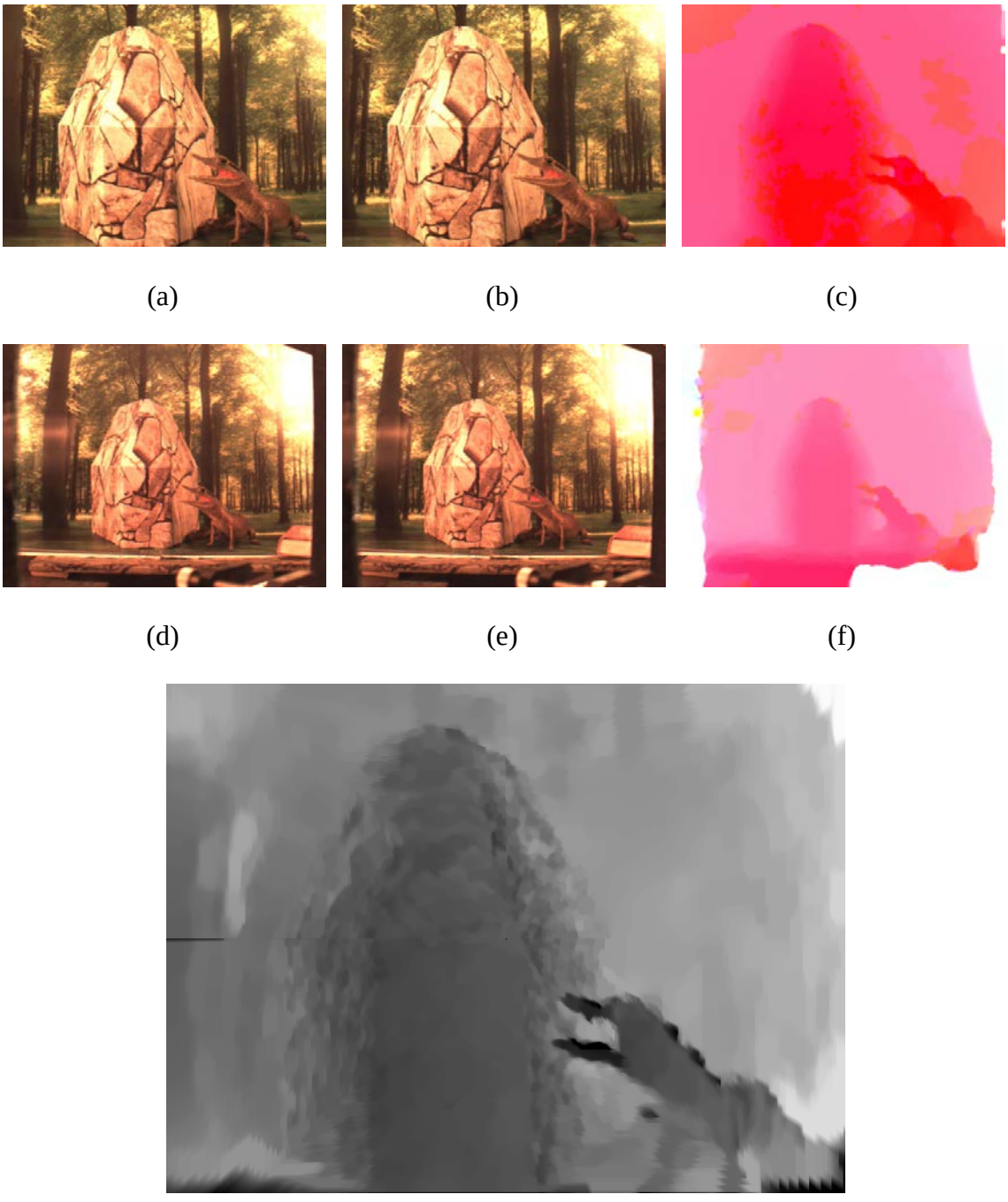


(a) (b) (c)

(d) (e) (f)

(g)

Figure 2.15. Flagstone image sequence, coaxial camera rig, graph cuts: (a) first front camera image, (b) second front camera image, (c) optical flow from front camera image pair, (d) first back camera image, (e) second back camera image, (f) optical flow from back camera image pair, (g) resulting depth map.



(a) (b) (c)

(d) (e) (f)

(g)

Figure 2.16. Flagstone with alligator image sequence, coaxial camera rig, graph cuts: (a) first front camera image, (b) second front camera image, (c) optical flow from front camera image pair, (d) first back camera image, (e) second back camera image, (f) optical flow from back camera image pair, (g) resulting depth map.

Table 2.2. Coaxial graph cuts: alignment errors, scene flow errors, and computational time.

	Fountain	Flagstone	Flagstone + alligator
ICPM Flow Alignment RMS Error	0.13 pixels	0.03 pixels	0.08 pixels
ICPM Scene flow error	3.0%	1.6%	3.6%
SfM Scene flow error	14.4%	4.0%	3.6%
ICPM Computational Time	74.6 seconds	59.2 seconds	64.5 seconds
SfM Computational Time	4.4 seconds	10.1 seconds	5.1 seconds

alignment. If the variational methods approach did not start with a reasonable estimate of Z (e.g., if the scene has numerous independently moving objects), the number of iterations required to reach a good solution could go up dramatically.

While initialization is an advantage in terms of computational speed for the variational methods solution, not having to do initialization is an advantage for the graph cuts solution. Graph cuts still requires a finite list of labels based on the environment, but this list has to cover only the desired working range of the camera rig and at the desired resolution of Z and $\dot{Z}$. Reducing the resolution of Z and $\dot{Z}$ can substantially decrease the computational speed.

Another advantage of the graph cuts solution is that graph cuts has a built-in stopping condition, which stops when the optimal solution is found for the given set of finite labels. This characteristic of graph cuts prevents the issues we see with the variational methods approach where determining the ideal stopping condition is dependent on the scene and the quality of the optical flow fields.

Another advantage of the graph cuts solution is that we can solve for Z and $\dot{Z}$ simultaneously. This advantage reduces errors that may result from lagging the solution

for $\dot{Z}$.

Graph cuts is a discrete methodology. This characteristic results in larger flow field alignment errors than the variational methods approach. Where the objective is to align the underlying images, variational methods have an advantage. Also related to the discrete approach of graph cuts is that the resolution of the depth map as well as the reconstruction of the scene flow will be discrete. This characteristic suggests that the optimal graph cuts label resolution should be selected to match the resolution of the optical flow algorithm being used.

CHAPTER 3

MULTIMODAL CAMERA RIG

3.1 Introduction

In computer vision, finding correspondences between rectified stereo image pairs is one of the most active research areas. Corresponding points in image pairs are typically found using pixel intensities, image features, or sometimes a combination of the two methods. Dense correspondences produce dense disparity maps, which can be used to estimate dense depth maps using the camera rig geometry. Additionally, corresponding points can be used to warp one image, the sensed image, into the second image, the reference image.

There is, however, a particular type of camera rig, the multimodal camera, where image-feature or pixel-intensity-based correspondence-finding algorithms do not work well or do not work at all. This result is due to image features or pixel intensities not having the same visual appearance when imaged at different wavelengths of light. In this chapter we provide background on the current state of the art in correspondence finding for multimodal camera rigs and then present a method of finding correspondences in pairs of multimodal image sequences using the perceived motion in the images instead of intercamera image features and/or pixel intensities.

We derive the relationship between the flow fields for a particular type of multimodal

camera rig, one where the two optical systems have two different magnifications. Like the binocular stereo rig, the cameras in this system have parallel optical axes, but unlike a traditional binocular stereo rig, the two optical system have two different magnifications. This difference in magnification produces optical flow fields whose ratio is a function of the distance to the scene. The scaling of the optical flow aids in the correspondence finding, allowing the degenerate case of frontal planar regions to be aligned. The equations for a system with different magnification optical systems can be applied to a standard binocular stereo rig by using the same focal lengths and same Z distances in the left and right cameras. As long as the scene does not contain frontal planar regions (a degenerate case), the method works equally well with rectified images from a standard binocular stereo rig.

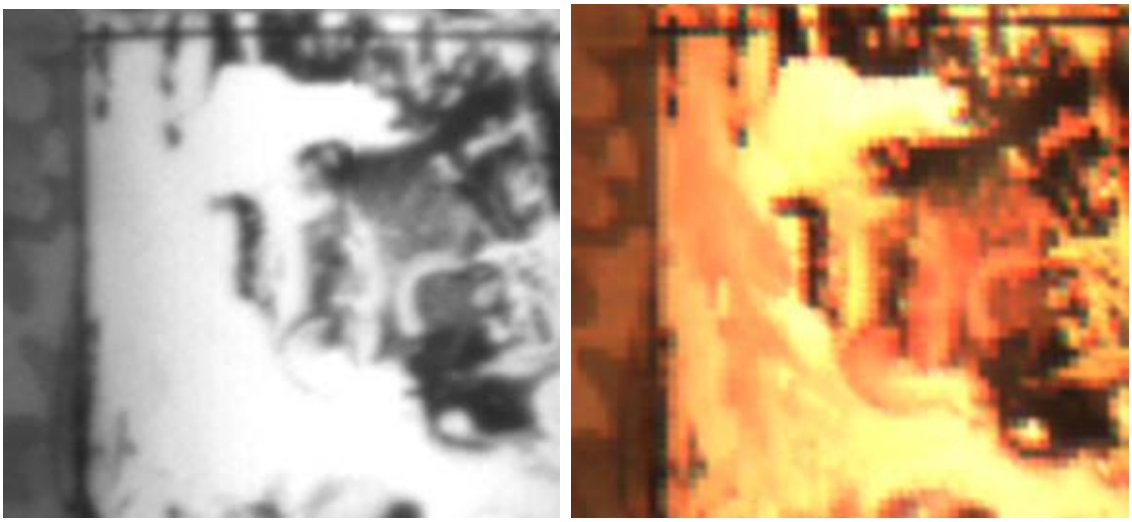
Using the relationship between the flow fields for a multimodal camera rig, we construct an energy minimization functional that when solved results in aligned flow fields. We present two numerical solutions to the energy minimization problem: a variational methods approach and one using graph cuts. We test the method on synthetic optical flow images and on three real-world scenes and present the resulting depth maps and accuracy metrics. We compare the accuracy of our results to that of the state-of-the-art multimodal methodology.

3.2 Related Work — Multimodal Camera Rigs

Aligning images from stereo rigs consisting of cameras with multimodal sensors has been an active research area for the last decade and a half. Initially inspired by the work done to match medical images to models [48], it has more recently been motivated by the

need for surveillance systems that use a combination of visible light and infrared (IR) cameras to detect targets. As noted by Yaman and Kalkan [49], traditional image alignment techniques used in stereo vision are not applicable to multimodal camera rigs because the pixel intensities can be substantially different in a visible light image vs. an IR image. This characteristic can be seen in Figure 3.1, which is an image pair taken with the IR/RGB multimodal camera rig used in our research.

Solutions to the multimodal problem currently fall into several broad categories. The first uses mutual information (MI). MI was originally proposed by Viola and Wells [48] to match medical images to models. Egnal [50] is reported to be the first to have used MI as a similarity measure to match multimodal stereo images. Since then, numerous improvements have been made including adaptive windowing [51], incorporating prior probabilities [52], regions of interest [53]-[55], and extending MI using gradient information [56]. According to Krotosky and Trivedi [38] "Due to high differences in



(a) (b)

Figure 3.1. Multimodal camera rig image pair: (a) IR and (b) RGB.

imaging characteristics, it is very difficult [to] find correspondences for the entire scene." No existing MI method has been reported to produce dense depth maps when used with RGB-IR image pairs.

More recently, local self-similarity (LSS), originally used in template matching, was proposed for use in a multimodal camera rig [57]. LSS is also a sparse technique.

The state of the art in multimodal stereo correspondence-finding technique uses a combination of scale-invariant feature transform (SIFT) and edge oriented histograms (EOH). In Auguilera et al. [58], points of interest are first found using a SIFT-based scale space representation. EOH descriptors are then used to further characterize the points of interest. Lastly, points of interest between the images in the multimodal image pairs are matched using the descriptor information.

The method we present avoids using visual similarity measures between the images from the two sensor types by computing the optical flow fields from the two sensors and then aligning the flow fields. This approach permits images with no common features to be aligned as long as there is motion between the camera and the scene, and the scene has enough texture at the different light wavelengths being imaged to produce optical flow.

Verri and Poggio [59] have shown that in many cases optical flow is not equivalent to the motion field. Optical flow algorithms have improved substantially since the Verri and Poggio paper (see [24], [31] for summaries of the progression of optical flow algorithm development), but optical flow errors caused by the aperture problem, non-Lambertian surfaces, and nonuniform or changing illumination still exist.

For finding image correspondences, however, the optical flow fields do not need to be equivalent to the motion fields. For example, errors caused by the aperture problem,

where only the motion tangential to edges is detected or errors caused by moving shadows, will be perceived by the two sensors identically and alignment is unaffected. The primary requirement is that the optical flow computation be invariant to different light wavelengths. To be invariant to different light wavelengths requires that the scene have visual texture perceptible under each wavelength of light being imaged. Given that there is sufficient visual texture at each wavelength being imaged and subject to the known deficiencies of optical flow computation mentioned above, the equations that govern the projection of the 3D scene onto the 2D image plane produce the same optical flow fields independent of the wavelength of light being imaged.

3.3 Energy Formulation

Referring to Figure 3.2, let $\bar{x}_l = (x_l, y_l)^T$ and $\bar{x}_r = (x_r, y_r)^T$ represent points in the image domain of the left and right cameras. Let $\bar{h}(\bar{x})$ be the disparity between $\bar{x}_l$ and $\bar{x}_r$ such that $\bar{x}_l$ and $\bar{x}_r + \bar{h}(\overline{\bar{x}_r})$ represent the same point $\bar{X}(\bar{x}_l) = (X, Y)$ in the scene. Let f_l and f_r be the focal lengths of the left and right cameras, respectively, and $Z_{l0}(\bar{x}_l)$ and $Z_{l1}(\bar{x}_l)$ be the distance between the optical center of the left camera and a point in the scene corresponding to $\bar{x}_l$ at time $t = 0$ and $t = 1$, the distance being measured along the optical axis. The difference along the Z axis for each point between $t = 0$ and $t = 1$ is $\Delta Z(\bar{x}_l)$. Let $\bar{X}$ be the distance from the optical axis to a point in the scene and $\Delta\bar{X}$ be the change in the distance from the optical axis between time $t = 0$ and $t = 1$. Let b be the stereo baseline. Let $\bar{w}_l$ and $\bar{w}_r$ be the projection of the 3D motion (the ideal optical flow) of a point in the scene onto the image planes of the left and right cameras, respectively.

We first derive equations for $\bar{h}(\bar{x}_l) = \begin{bmatrix} h_x(x_l) \\ h_y(y_l) \end{bmatrix}$, which is the disparity in x and y with

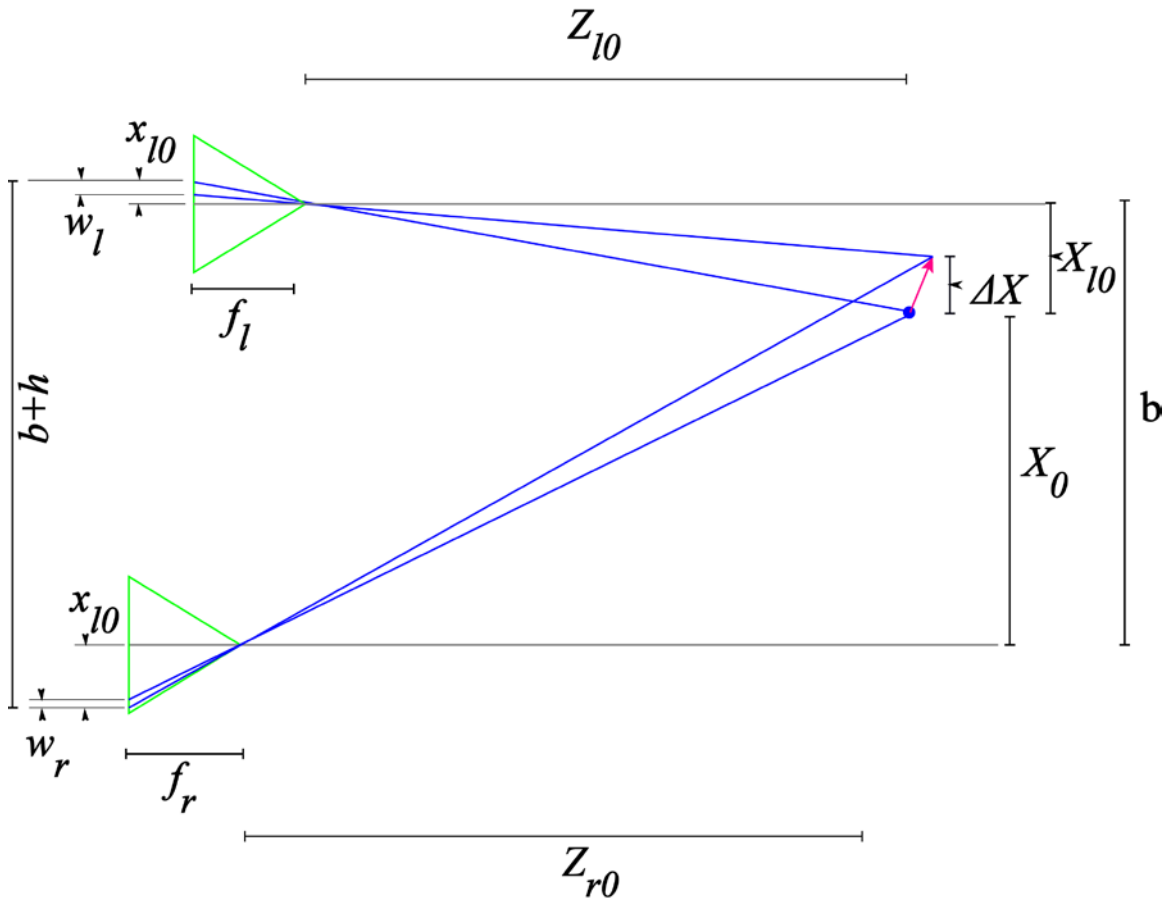


Figure 3.2 Multimodal stereo camera rig geometry X-Z view.

the left image being the reference image. For the x direction, we start with the projection equations for a pinhole camera

$$x_l = -\frac{f_l X_l}{Z_l} \tag{3.1}$$

$$x_r = -\frac{f_r X_r}{Z_r} \tag{3.2}$$

where

$$b = X_r - X_l \tag{3.3}$$

is the stereo baseline. Solving for the disparity in the x direction gives

$$x_l - x_r = h_x = \frac{f_r X_r}{Z_r} - \frac{f_l X_l}{Z_l}. \tag{3.4}$$

Reducing gives

$$h_x = \frac{\left(-\frac{f_r}{f_l} x_l Z_l\right) + f_r b + x_l Z_l - x_l d}{Z_l + d} \tag{3.5}$$

where $d = Z_l - Z_r$ is the difference in Z distance between the optical centers of the left camera and the right camera.

If the focal lengths in the left and right cameras are equal (i.e., $d = 0$ and $f_l = f_r$), (3.5) reduces to the well-known binocular stereo disparity equation

$$h = \frac{fb}{Z}. \tag{3.6}$$

Referring to Figure 3.3, we use the same method to derive the disparity in the y direction to arrive at

$$h_y = \frac{y_l\left(Z_l + d - \frac{f_r}{f_l} Z_l\right)}{Z_l + d}. \tag{3.7}$$

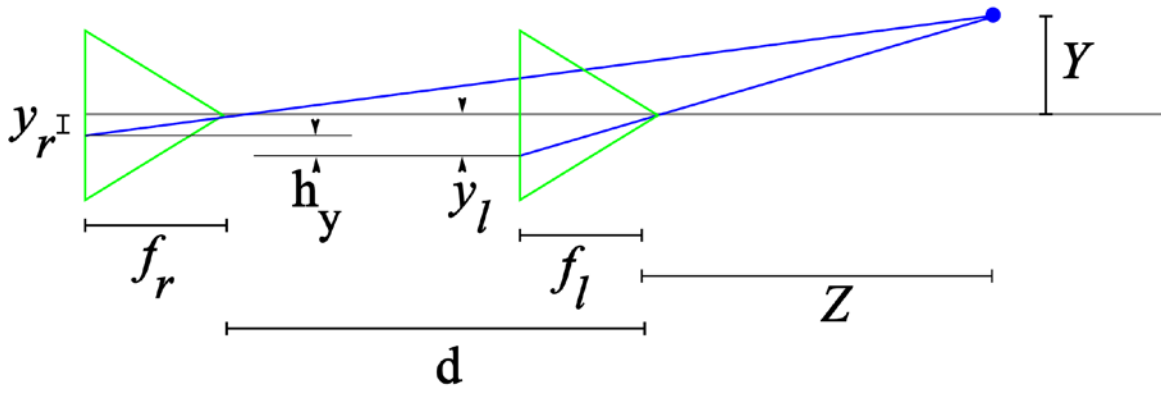


Figure 3.3. Multimodal stereo camera rig geometry Y-Z view.

This equation reduces to the equation for a traditional stereo epipolar line in a rectified image pair for $d = 0$ and $f_l = f_r$

$$h_y(x_f) = 0. \tag{3.8}$$

We now find the relationship between the optical flow, which depends on both Z and ΔZ. Because the derivation is done using continuous derivatives, we use $\dot{Z}$ instead of ΔZ, but when we move back to a discrete formulation we will replace $\dot{Z}$ with ΔZ. Once again, we start with the projection equations and take the derivative with respect to time

$$\frac{dx}{dt} = w_x = -f\frac{d}{dt}\left(\frac{X}{Z}\right) \tag{3.9}$$

$$\frac{dy}{dt} = w_y = -f\frac{d}{dt}\left(\frac{Y}{Z}\right) \tag{3.10}$$

$$w_x = \frac{x\dot{Z} - f\dot{X}}{Z} \tag{3.11}$$

$$w_y = \frac{y\dot{Z} - f\dot{Y}}{Z} \tag{3.12}$$

which can be written in homogeneous coordinates as

$$\bar{P} = \begin{bmatrix} 1 & 0 & x/f & 0 \\ 0 & 1 & y/f & 0 \\ 0 & 0 & 0 & -Z/f \end{bmatrix} \tag{3.13}$$

$$\bar{w} = \begin{bmatrix} 1 & 0 & -x/f & 0 \\ 0 & 1 & -y/f & 0 \\ 0 & 0 & 0 & -Z/f \end{bmatrix} \begin{bmatrix} \dot{X} \\ \dot{Y} \\ \dot{Z} \\ 1 \end{bmatrix} = \begin{bmatrix} \dot{X} - \frac{x\dot{Z}}{f} \\ \dot{Y} - \frac{y\dot{Z}}{f} \\ -\frac{Z}{f} \end{bmatrix} = -\begin{bmatrix} \frac{f\dot{X} - x\dot{Z}}{Z} \\ \frac{f\dot{Y} - y\dot{Z}}{Z} \end{bmatrix}. \tag{3.14}$$

Adding image frame timing to (3.11) and (3.12) gives

$$\bar{w}_l = \frac{x_{l0}\dot{Z} - f_l\dot{\bar{X}}}{Z_{l1}} \tag{3.15}$$

$$\bar{w}_r = \frac{x_{r0}\dot{Z} - f_r\dot{\bar{X}}}{Z_{r1}} \tag{3.16}$$

for the left and right cameras. Solving for $\dot{\bar{X}}$ and setting the resulting equations equal to each other gives

$$p(\bar{x}_l)\bar{w}_l(\bar{x}_l) = \bar{g}(\bar{x}_f)\bar{w}_r\left(\bar{x}_l + \bar{h}(\bar{x}_l)\right) \tag{3.17}$$

where

$$p(\bar{x}_l) = \left(\frac{f_l}{f_r}\right)\left(\frac{Z_{r1}}{Z_{l1}}\right) \tag{3.18}$$

and

$$\bar{g}(\bar{x}_l) = \left(\frac{f_l Z_{r1} \bar{w}_r}{f_l Z_{r1} \bar{w}_{r+} f_l \bar{x}_{r0} Z - f_r \bar{x}_{l0} \dot{Z}}\right), \tag{3.19}$$

which can be written as an energy functional

$$E_{match} = \sum_{p \in \mathcal{P}} \left[p(\bar{x}_l)\bar{w}_l(\bar{x}_l) - \bar{g}(\bar{x}_l)\bar{w}_r\left(\bar{x}_l + \bar{h}(\bar{x}_l)\right)\right]^2 \tag{3.20}$$

$$E_{smooth} = \sum_{p \in \mathcal{P}} \|\nabla Z_l(\bar{x}_l)\|^2 \tag{3.21}$$

$$E_{total} = \gamma E_{match} + \bar{\alpha} E_{smooth} \tag{3.22}$$

where

$$\bar{\alpha} = \left(\alpha_x, \alpha_y\right)^T. \tag{3.23}$$

3.4 Numerical Solutions

Equation (3.22) has the same form as (2.18) and can be solved using variational methods or graph cuts in a similar manner to the solutions presented in Sections 2.5 and 2.6 of this dissertation. In this section we summarize the differences between the solutions for the multimodal stereo vs. the coaxial camera rig.

3.5 Variational Methods

The variational methods approach requires that the energy be expressed in a continuous form such that the first variation can be found. Additionally, we separate variations in optical flow due to Z from variations in optical flow due to $\dot{Z}$ in the formulation to streamline the gradient descent computation.

3.5.1 Euler-Lagrange

We rewrite (3.20) and (3.21) in continuous form and we reexpress the smoothing term using an L2 norm

$$E_{match} = \frac{1}{2}\int_a^b \left[p(\bar{x}_l)\bar{w}_l(\bar{x}_l) - \bar{g}(\bar{x}_l)\bar{w}_r\left(\bar{x}_l + \bar{h}(\bar{x}_l)\right)\right]^2 d\bar{x} \tag{3.24}$$

$$E_{smooth\_Z} = \frac{1}{2}\int_a^b \|\nabla Z_l(\bar{x}_l)\|^2 \, d\bar{x} \tag{3.25}$$

where

$$p(\bar{x}_l) = \left(\frac{f_r}{f_l}\right)\left(\frac{Z_{l1}}{Z_{r1}}\right) \tag{3.26}$$

$$\bar{g}(\bar{x}_l) = \left(\frac{f_l Z_{r1}\bar{w}_r}{f_l Z_{r1}\bar{w}_r + f_l\bar{x}_{r0}\dot{Z} - f_r\bar{x}_{l0}\dot{Z}}\right). \tag{3.27}$$

We can now take the first variation of equations (3.24) and (3.25) with respect to Z

$$\begin{aligned}\gamma w_z(p'w_l + pw_l' - \bar{g}'w_r\left(\bar{x}_l + \bar{h}(\bar{x}_l)\right) \\ -\bar{g}w_r'\left(\bar{x}_l + \bar{h}(\bar{x}_l)\right)\bar{h}') - \bar{\alpha}\nabla^2 Z_1\end{aligned} \tag{3.28}$$

where

$$h_x{}' = \frac{\partial h_x}{\partial Z} = \frac{-\frac{f_r}{f_l}x_l + x_l}{Z_l + d} - \frac{\left(-\frac{f_r}{f_l}x_l Z_l\right) + f_r b + x_l Z_l - x_l d}{(Z_l + d)^2} \tag{3.29}$$

$$h_y{}' = \frac{\partial h_y}{\partial Z} = \frac{y_l\left(1 - \frac{f_r}{f_l}\right)}{Z_l + d} - h_y = \frac{y_l\left(Z_l + d - \frac{f_r}{f_l}Z_l\right)}{(Z_l + d)^2} \tag{3.30}$$

$$p' = \frac{\partial p}{\partial Z} = \left(\frac{f_l}{f_r}\right)\left(\frac{1}{Z_{l1}} + \frac{Z_{r1}}{(Z_{l1})^2}\right) \tag{3.31}$$

$$w_l{}' = \frac{\partial w_l}{\partial Z} = -\frac{w_l}{Z_l} \tag{3.32}$$

$$w_r{}' = \frac{\partial w_r}{\partial Z} = -\frac{w_r}{Z_r} \tag{3.33}$$

$$g' = \frac{\partial g}{\partial Z} = \frac{f_l w_r + f_l Z_{r1} w_r'}{f_l Z_{r1} w_r + f_l x_{r0}\dot{Z} - f_r x_{l0}\dot{Z}} + \frac{(f_l Z_{r1} w_r)(f_l w_r + f_l Z_{r1} w_r')}{(f_l Z_{r1} w_r + f_l x_{r0}\dot{Z} - f_r x_{l0}\dot{Z})^2} \tag{3.34}$$

$$\bar{\alpha}\nabla^2 Z_{l1} = \alpha_x \frac{\partial^2 Z_{l1}}{\partial x^2} + \alpha_y \frac{\partial^2 Z_{l1}}{\partial y^2} \tag{3.35}$$

$$w_z = p(\bar{x}_l) w_l(\bar{x}_l) - \bar{g}(\bar{x}_l) w_l\left(\bar{x}_l + \bar{h}(\bar{x}_l)\right). \tag{3.36}$$

The Euler-Lagrange equations (one for the *x* direction and the other for the *y* direction) are solved using the gradient descent method.

3.5.2 Implementation Details

3.5.2.1 Discrete Laplacian

The discrete Laplacian is computed using a finite difference scheme.

3.5.2.2 Initialization

We initialize the value of Z by taking the optical flow in the center pixel of the left (IR) image and estimate the scaled optical flow and disparity for $Z = \{1, 2, 3, \dots\}$ that should be perceived by the right camera based on the camera rig geometry. When the estimated disparity and optical flow intersect with the actual disparity and optical flow value from the optical flow field computed from images from the right camera, we have an estimate of the depth at that point. Using this estimate of depth at one location, we estimate the $\dot{X}$ velocity. We then estimate Z at all points using $\dot{X}$. The Z estimate will contain errors in many if not most locations for a number of reasons, but this method produces a usable initial estimate.

3.5.2.3 Resampling to a Discrete Grid

Like the coaxial camera rig, the gradient descent results in a new estimate of Z at $t = n + 1$ after each step. This estimate, being offset spatially by the optical flow, must be resampled onto the pixel grid.

3.5.2.4 Stopping Criteria

We used the same two stopping criteria as with the coaxial camera formulation, depending on the quality of the flow fields and the value chosen for $\bar{\alpha}$. When the flow fields closely represent the motion fields and $\bar{\alpha}$ is small (minimal Z smoothing), we compute

$$error_{flow\ match} = \left[p(\bar{x}_l)\bar{w}_l(\bar{x}_l) - \bar{g}(\bar{x}_l)\bar{w}_r\left(\bar{x}_l + \bar{h}(\bar{x}_l)\right)\right]^2 \tag{3.37}$$

after each step in the gradient descent. Equation (3.37) is a measure of the mismatch in registration of the two flow fields. We stop iterating when (3.37) falls below a predetermined value. We used 0.01 pixels as the threshold, the same as with the coaxial camera rig.

Where the flow fields are noisy, it is necessary to increase $\bar{\alpha}$ to get good results. With more substantial smoothing, the smoothing term (3.25) appears to pull the Z estimate away from the correct value if γ is large and/or if many iterations are performed. This result is particularly evident around discontinuities in the scene, which are worse for the multimodal stereo rig, than for the coaxial camera rig. In this case we stopped the iterations when the smoothing term (3.25) was approximately equal to, but of opposite sign to, the matching term (3.24). This latter approach produced larger residual values of w_z, but the experiments show that it results in more accurate depth estimations near discontinuities in the scene.

3.5.2.5 Algorithm

1) Compute $\bar{w}_l$ and $\bar{w}_r$.
2) Smooth $\bar{w}_l$ and $\bar{w}_r$.
3) Initialize Z.
4) Iterate until stopping condition met.
    a) For each epipolar line:
        i) update Z estimate for one gradient descent step,
        ii) resample Z estimate to grid,
        iii) compute $\dot{z}$,

iv) update $g(\bar{x}_l)$.

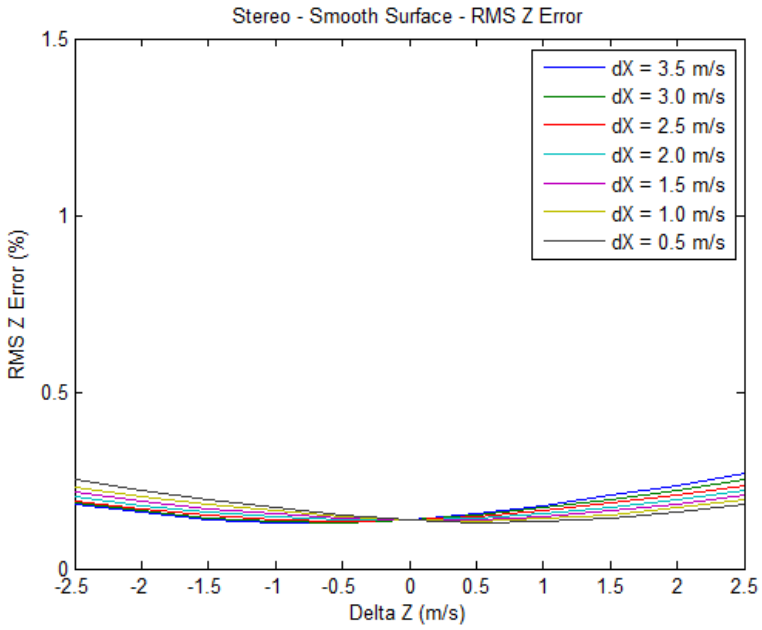
3.5.3 Experimental Results

As with the coaxial camera rig, we tested the multimodal camera rig method on both synthetic optical flow fields and on real image sequences. The purpose of using the synthetic optical flow fields was to verify that the energy formulation, when solved, resulted in alignment of the underlying images and in accurate depth estimations. With optical flow fields that are an accurate projection of the motion field, the reconstructed depth map in all nonoccluded areas will line up with the ground truth to within the numerical estimation error.

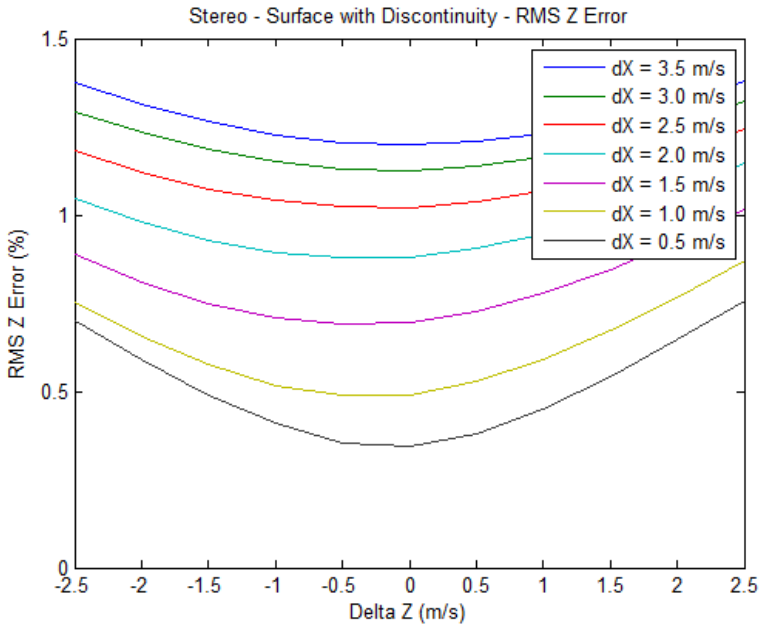
3.5.3.1 Synthetic Optical Flow Fields

For the synthetic optical flow field experiments, we used the same simulated optical flow field as for the coaxial camera. The scene used to generate the synthetic optical flow field did not have frontal planar regions in it and thus worked equally well with a binocular stereo camera geometry. To determine the accuracy of the resulting image alignment, we reconstructed the depth map along a horizontal epipolar line using the results of registration and compared the reconstructed depth map with the original scene geometry computing both the RMS disparity error and the resulting RMS depth error.

Figures 3.4(a) and 3.5(a) show the results for a smooth scene without any occlusions. The worst-case RMS depth error is < 0.25% and worst-case RMS disparity errors < 0.01 pixels. The accuracy is slightly reduced as $\dot{Z}$ increases and $\dot{X}$ decreases. As with the coaxial camera rig, we believe that the increased error is due to lagging the solution for $\dot{Z}$.

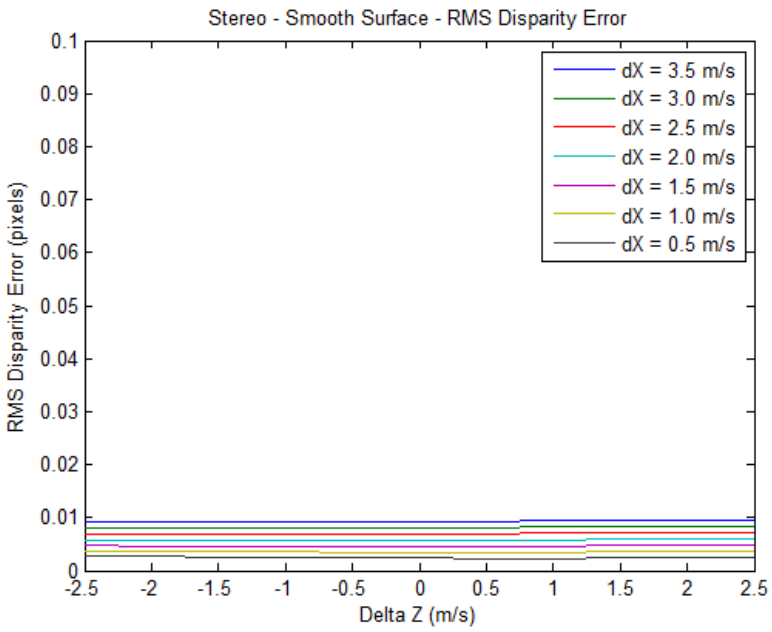


(a)

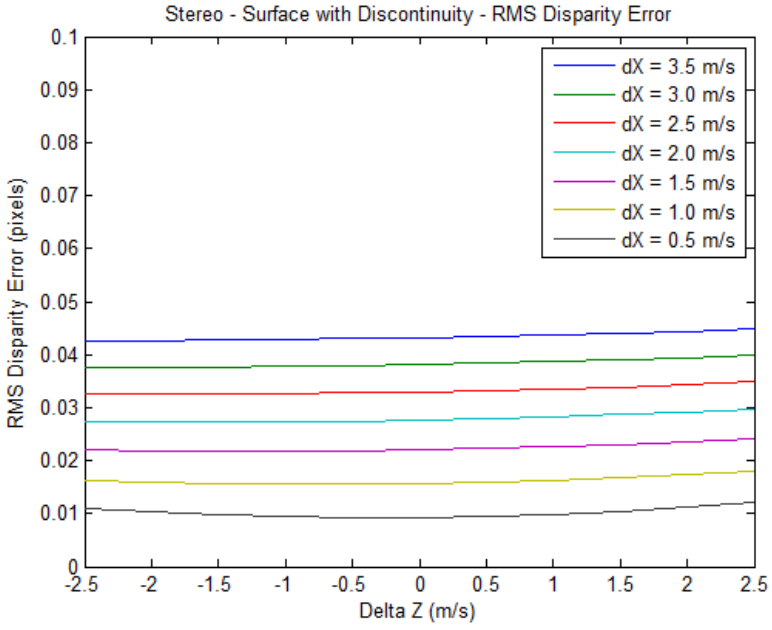


(b)

Figure 3.4. RMS Z error for multimodal stereo camera rig using synthetic flow fields. (a) Smooth surface. (b) Surface with discontinuities and occlusions.



(a)



(b)

Figure 3.5. RMS disparity error error for multimodal stereo camera rig using synthetic flow fields. (a) Smooth surface. (b) Surface with discontinuities and occlusions.

Figures 3.4(b) and 3.5(b) show the results for a scene with a large occlusion caused by a large (8 m) discontinuity in the simulated scene. The RMS error increases between a smooth scene and an occluded scene is similar to that of the coaxial camera rig.

3.5.3.2 Flow Fields From Camera Images

The multimodal stereo camera rig consists of one camera with an RGB sensor and a second camera that is sensitive only to IR light above 700 nm (Figure 3.6). The RGB camera is a Point Gray 0.3MP Color Firefly MV 1/3" CMOS computer vision camera with global shutter. The IR camera is a Point Gray 1.3MP Monochrome Flea3 1/2" CMOS computer vision camera with global shutter. The Flea3 uses an On

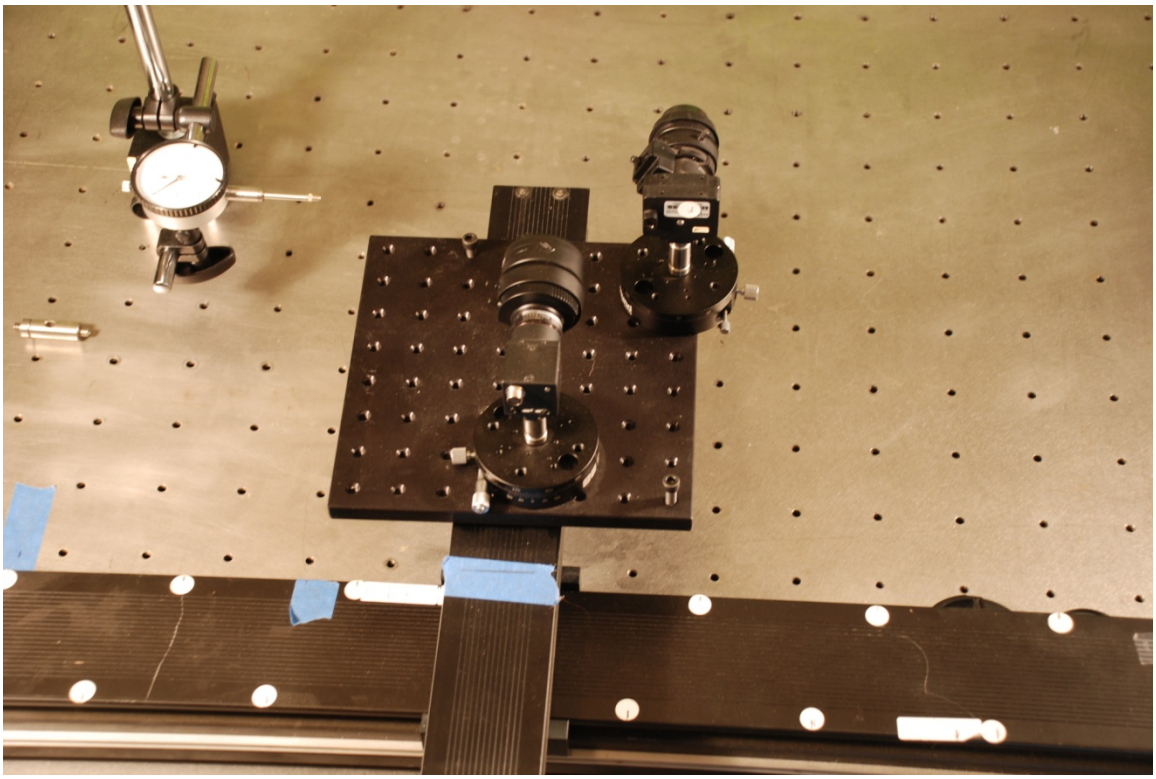


Figure 3.6. Multimodal camera rig on XY table.

Semiconductor CMOS sensor that is sensitive to light wavelengths from 300 nm to approximately 1100 nm. It does not come with an internal IR filter. We added an Edmund Optics part number 64887 UV/VIS cut-off filter that blocks UV and visible light below 700 nm, which effectively converts the Flea3 into a 700 nm to 1100 nm near-IR (NIR) camera.

The camera rig was mounted on the same precision XY table as the coaxial camera rig (section 2.5.3.2). The same scenes were used for the multimodal camera rig as for the coaxial camera rig. The cameras in the multimodal stereo rig had 4.8 micron (IR) and 6 micron (RGB) square pixels and approximately 3.8 mm (IR) and 8.0 mm (RGB) focal lengths. The cameras were calibrated using Cal Tech's Camera Calibration Toolbox [41] based on the work of Zhang et al. [42], [43]. The Flea3 was calibrated with the IR pass filter in place.

The scenes are shown again in Figures 3.7, 3.8, and 3.9 (a), (b), (d), and (e) and the resulting optical flow in (c) and (f). The camera rig was translated 20 mm between image frames, which equates to a velocity of 0.6 m/s for a 30 fps frame rate. We set $\gamma = 2 \cdot 10^6$ and $\alpha = [.05, .01]$. We used the large-scale optical flow algorithm from Brox and Malik [24].

Accuracy of flow field alignment was measured by warping the left camera (IR) flow field based on the estimated depth map and taking the RMS error between the warped left flow field and the flow field obtained from the right camera. Figure 3.10 shows how this is done along one horizontal line for an RGB/IR image pair of the flagstone scene.

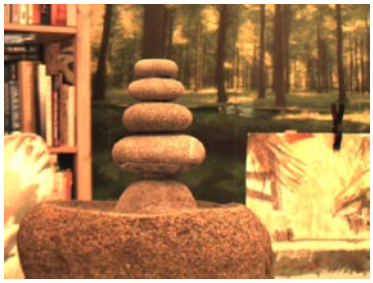
We compare our method, ICPM, with the state-of-the-art multimodal method, SIFT-EOH, from Aguilera et al. [58]. Figure 3.11 shows the sparse SIFT-EOH matches for

(a)

(b)

(c)

(d)

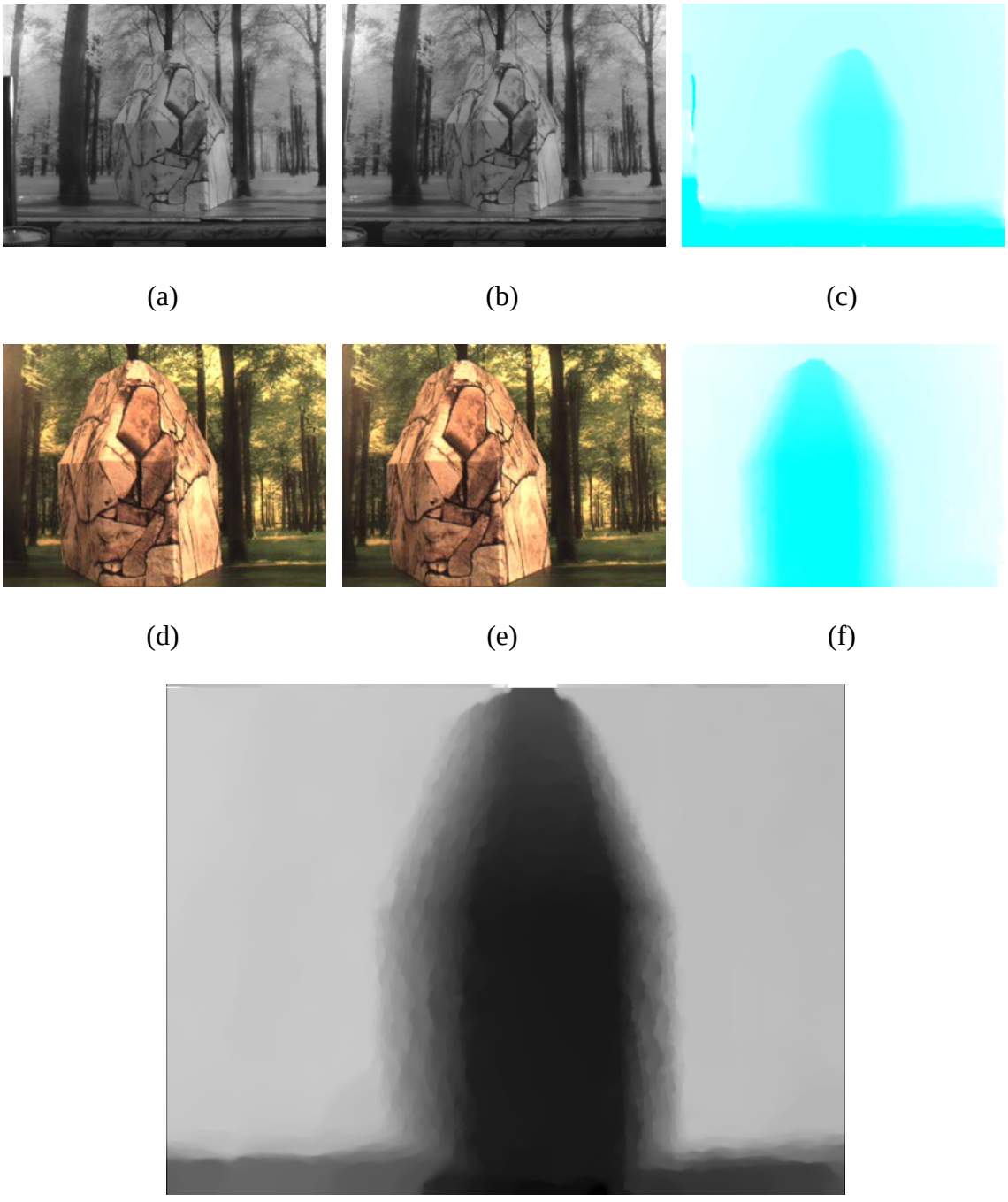


(e)

(f)

(g)

Figure 3.7. Fountain image sequence, multimodal stereo rig, variational methods: (a) first IR image, (b) second IR image, (c) optical flow from IR image pair, (d) first RGB image, (e) second RGB image, (f) optical flow from RGB image pair, (g) resulting depth map.

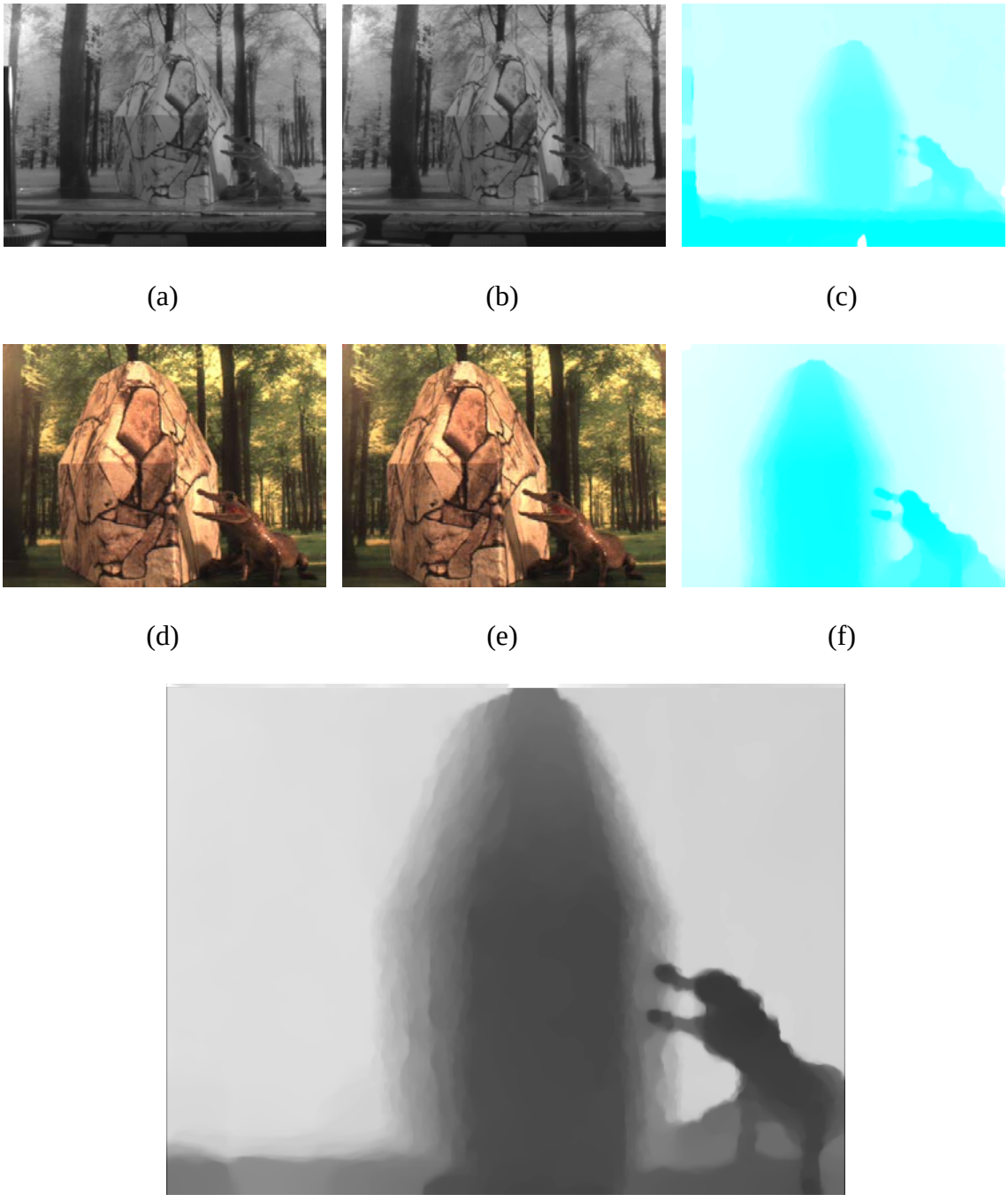


(a) (b) (c)

(d) (e) (f)

(g)

Figure 3.8. Flagstone image sequence, multimodal stereo rig, variational methods: (a) first IR image, (b) second IR image, (c) optical flow from IR image pair, (d) first RGB image, (e) second RGB image, (f) optical flow from RGB image pair, (g) resulting depth map.



(a) (b) (c)

(d) (e) (f)

(g)

Figure 3.9. Flagstone with alligator image sequence, multimodal stereo rig, variational methods: (a) first IR image, (b) second IR image, (c) optical flow from IR image pair, (d) first RGB image, (e) second RGB image, (f) optical flow from RGB image pair, (g) resulting depth map.

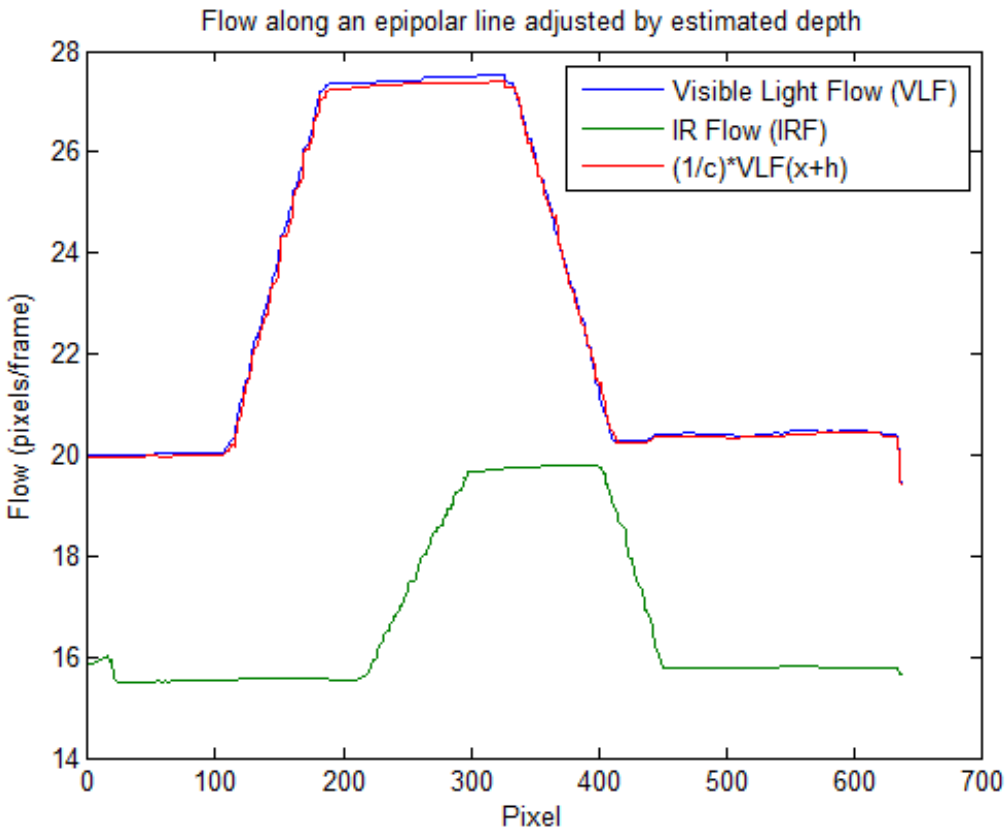
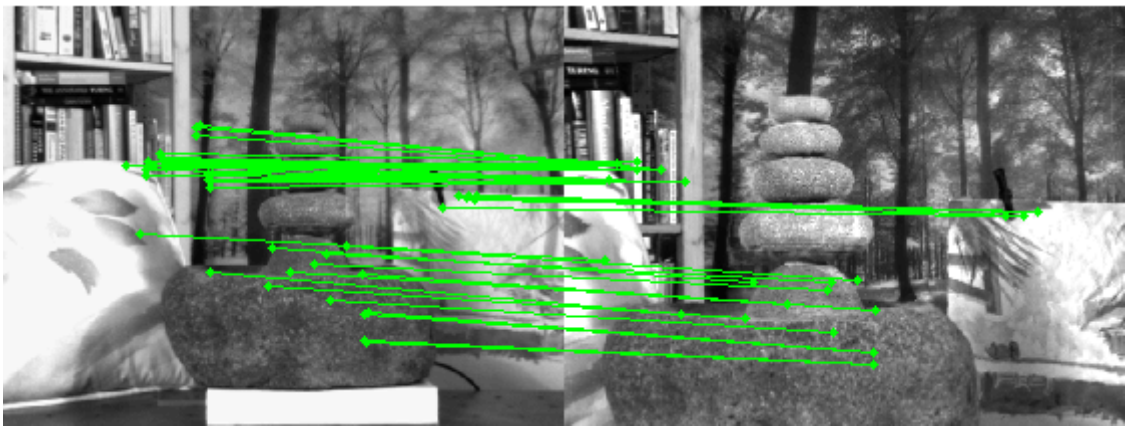


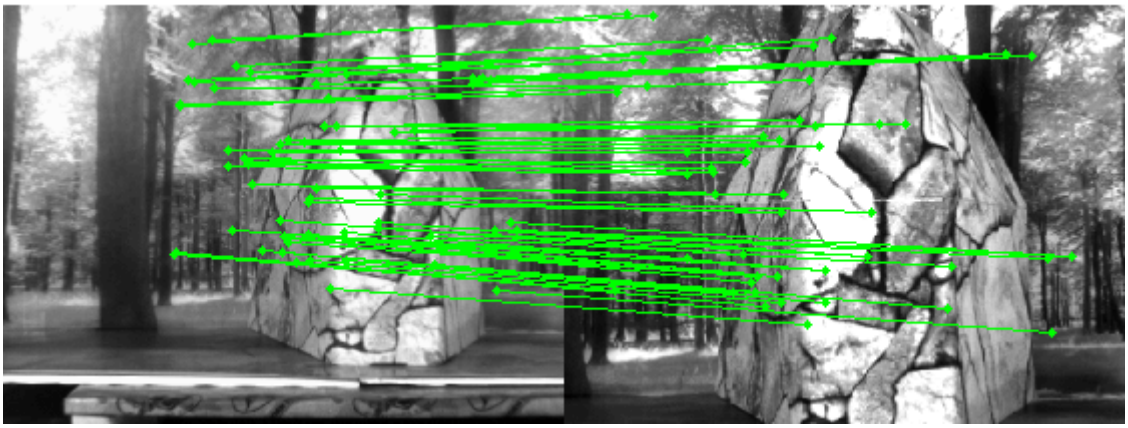
Figure 3.10. IR flow warped to match RGB flow using the estimated depth map for an epipolar line.

each of the three scenes and Table 3.1 shows the flow field alignment errors in pixels for each scene as well as the accuracy of the scene flow using ICPM and SIFT-EOH. The dense ICPM scene flow error includes all nonoccluded pixels, whereas the sparse ICPM scene error uses only the correspondences found using the SIFT-EOH method. To the first place after the decimal point, the scene flow error for ICPM is the same whether dense estimation or sparse estimation is used.

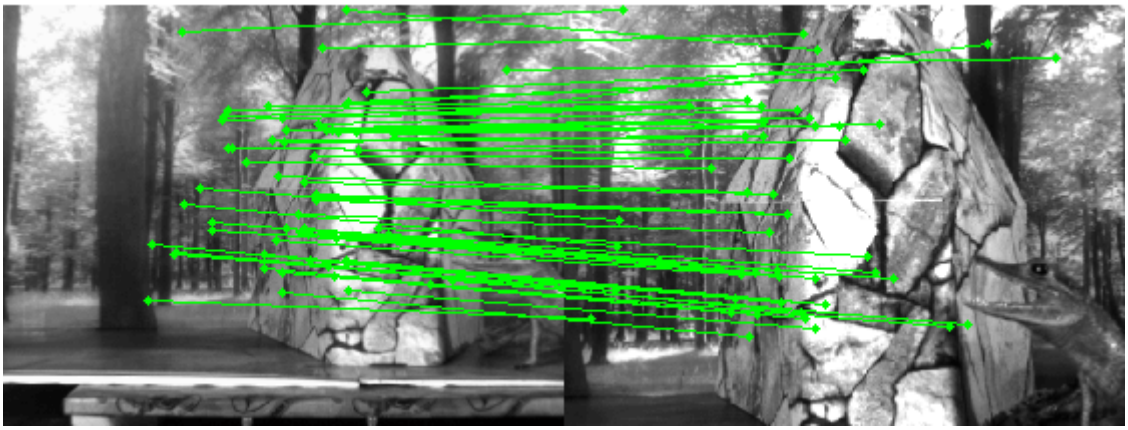
Depending on the scene, ICPM shows a reduction in the error in the scene flow by 77% to 88% compared to SIFT-EOH. Additionally, ICPM produces dense depth maps and scene flow estimates whereas SIFT-EOH produces matches for less than 0.1% of the



(a)



(b)



(c)

Figure 3.11. SIFT-EOH correspondences. (a) Fountain scene. (b) Flagstone scene. (c) Flagstone with alligator scene.

Table 3.1. Multimodal variational methods: alignment errors, scene flow errors, and computational time.

	Fountain	Flagstone	Flagstone + alligator
ICPM Flow Alignment Error	0.12 pixels	0.09 pixels	0.10 pixels
ICPM Dense Scene flow error	4.2%	1.9%	1.1%
ICPM Sparse Scene flow error	4.2%	1.9%	1.1%
SIFT+EOH Scene Flow error	30.2%	8.6%	8.9%
ICPM Computational Time	42.1 seconds	45.5 seconds	40.5 seconds
SIFT+EOH Computational Time	290.4 seconds	323.5 seconds	540.2 seconds

pixels.

Figures 3.7(f), 3.8(f), and 3.9(f) show the dense depth maps. In the dense depth maps, the closer the object is to the camera, the darker the pixel. Table 3.1 shows the flow field alignment errors, scene flow errors, and computational time.

3.5.4 Discussion

The depth maps are a reasonably good visual representation of the 3D shape of the objects in the scene. The main difference between the results from the stereo rig and the results from the coaxial camera rig is that the occluded areas in the stereo rig are larger. For the variational methods approach, this characteristic is not as visually noticeable as for the graph cuts approach as we will see in the next section. This result is due to the greater amount of smoothing in the variational approach. The error in reconstructing occluded areas is not isolated to motion-based correspondence finding, but the same reconstruction error occurs for any multicamera rig where one camera cannot see a

portion of the image. For the variational methods algorithm, the result is that the smoothing term in the energy functional creates a depth gradient that joins either sides of the occluded area.

As noted previously, depending on the scene our method is substantially (77% to 89% reduction in scene error) more accurate than the state-of-the-art SIFT-EOH method.

Although SIFT-EOH is not directly matching pixel intensities, it does match features using visual characteristics. The greater the difference in visual appearance of detected features between the multimodal image pairs, the more difficulty any method based on of visual similarity will have. ICPM will match image pairs that do not have any visual similarity as long as both images produce optical that is a reasonable representation of the projected scene flow.

3.6 Graph Cuts

3.6.1 Implementation Details

The graph cuts implementation for the multimodal stereo rig is nearly identical to that of the coaxial camera rig. The only difference is that the costs for the multimodal camera rig are computed using (3.20) and (3.21) instead of (2.20) and (2.21).

3.6.2 Experimental Results

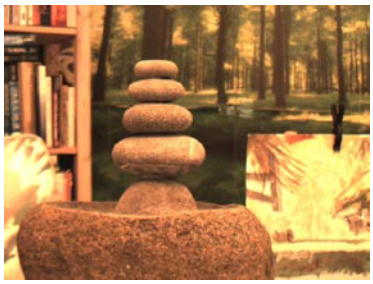
We tested the method on images from the three scenes used previously. The scenes are shown in Figures 3.12, 3.13, and 3.14(a)–(e) and the optical flow in (c) and (f). The resulting depth maps are shown in Figures 3.12(f), 3.13(f), and 3.14(f). The alignment accuracy, scene flow accuracy, and computational time are reported in Table 3.2.

(a)

(b)

(c)

(d)



(e)

(f)

(g)

Figure 3.12. Fountain image sequence, multimodal stereo rig, graph cuts: (a) first IR image, (b) second IR image, (c) optical flow from IR image pair, (d) first RGB image, (e) second RGB image, (f) optical flow from RGB image pair, (g) resulting depth map.

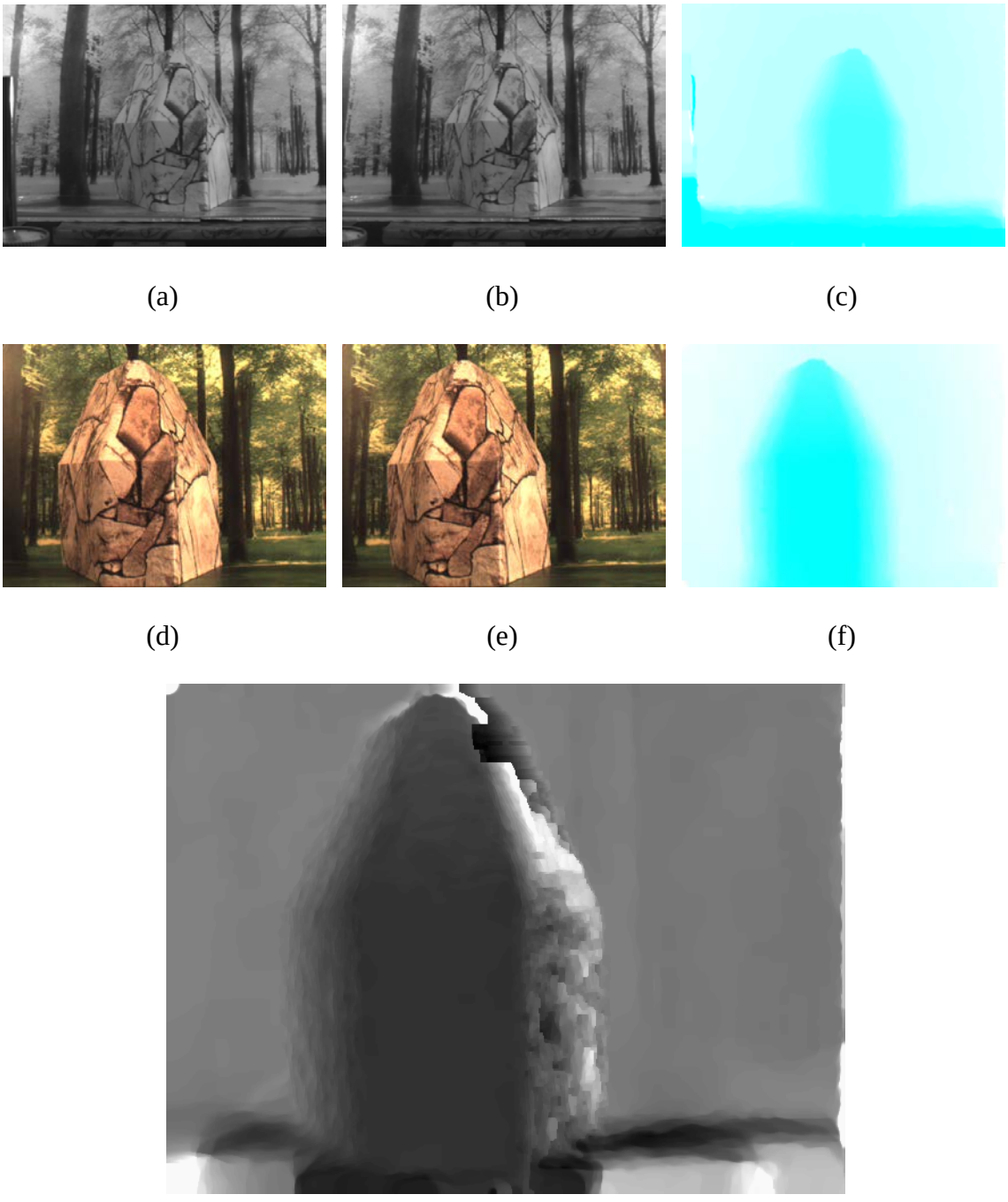


Figure 3.13. Flagstone image sequence, multimodal stereo rig, graph cuts: (a) first IR image, (b) second IR image, (c) optical flow from IR image pair, (d) first RGB image, (e) second RGB image, (f) optical flow from RGB image pair, (g) resulting depth map.

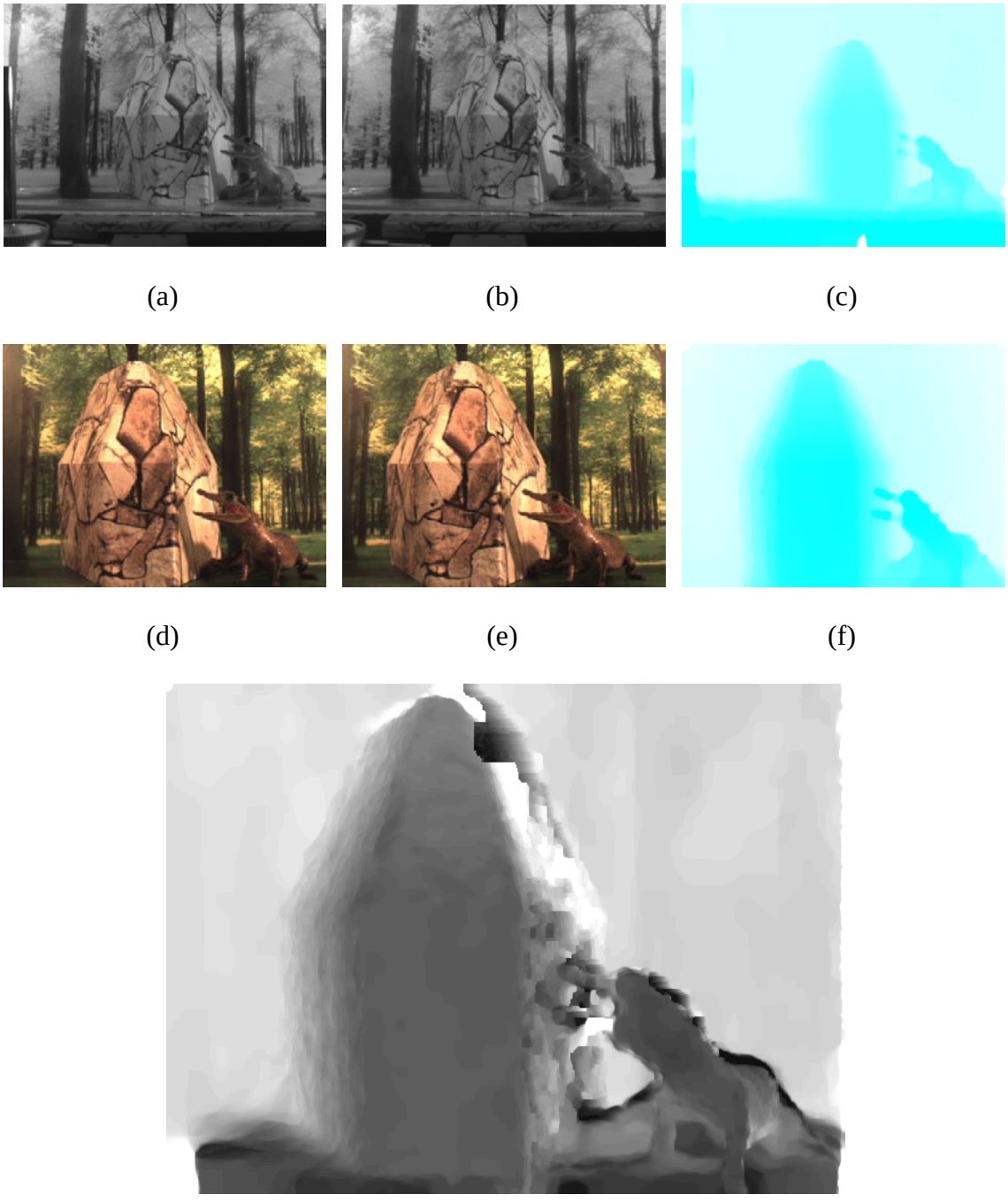


Figure 3.14. Flagstone with alligator image sequence, multimodal stereo rig, graph cuts: (a) first IR image, (b) second IR image, (c) optical flow from IR image pair, (d) first RGB image, (e) second RGB image, (f) optical flow from RGB image pair, (g) resulting depth map.

Table 3.2 Multimodal graph cuts methods: alignment errors, scene flow errors, and computational time.

	Fountain	Flagstone	Flagstone + alligator
ICPM Flow Alignment Error	0.17 pixels	0.12 pixels	0.13 pixels
ICPM Scene flow error	1.6%	2.2%	1.7%
SIFT+EOH scene flow error	30.2%	8.6%	8.9%
ICPM Computational Time	48.3 seconds	33.6 seconds	32.4 seconds
SIFT+EOH Computational time	290.4 seconds	323.5 seconds	540.2 seconds

3.6.3 Discussion

Visually, the depth maps are similar to those from the variational methods approach, but with some blockish features that are a common characteristic of graph cuts, particularly when using the L1 norm for regularization. As with the depth maps from the variational methods approach, there are also some visual deformities in the reconstructed depth maps due to the occlusions. However, the graph cuts solution responds to occluded areas differently. This difference is particularly noticeable in the fountain image where there is a 76-pixel occlusion between the left edge of the fountain and the background, and a 46-pixel occlusion between the right edge of the fountain and the background. The graph cuts solution produces a fairly significant smearing of the fountain in the area of the occlusion. Alignment errors are similar to those found using variational methods.

In general, the variational methods depth maps are smoother than the depth maps from graph cuts. The smoothness gives them a more visually pleasing look, but at the cost of losing some of the finer details. This visual characteristic is particularly evident in the flagstone-alligator image sequence where both the edges of the flagstone structure and the alligator's jaws are blended with the background. This result suggests that for

reconstruction for visualization purposes, variational methods might be the preferred methodology, but for reconstruction of fine details, the graph cuts solution might be preferable.

CHAPTER 4

CONCLUSIONS

In computer vision, finding correspondences between image pairs taken from different camera perspectives is one of the most active research areas. Corresponding points in image pairs are typically found using pixel intensities, image features, or some combination of the two.

There are, however, two types of camera rigs, the multimodal camera rig and the coaxial camera rig, where image-feature- or pixel-intensity-based correspondence-finding algorithms do not work well or do not work at all. For multimodal camera rigs, the reason for poor performance of traditional correspondence-finding algorithms is due to image features or pixel intensities not having the same visual appearance when imaged at different wavelengths of light. For coaxial camera rigs, failure of traditional correspondence-finding algorithms is due to the lack of disparity in the center region of image pairs. Both of these camera rigs have numerous uses if the images from the two cameras can be aligned in a way that image registration and 3D reconstruction were possible.

We have addressed the challenge of finding correspondences between pairs of image sequences where image features or pixel intensities do not work by finding correspondences using the motion in the scene. We have demonstrated the capability of

this technique by doing 3D reconstruction using image sequences taken with both a coaxial camera rig as well as a multimodal camera rig.

Motion-based correspondences provide an alternative to image-feature- or pixel-intensity-based methods for aligning images, but they produce a redundant set of constraints in image sequences that can be aligned using image features or pixel intensities for finding correspondences. Although not explored in this dissertation, this additional information could be useful when combined with existing correspondence-finding techniques, in particular for improving scene flow estimation as well as for improvements to the coaxial camera in the outer regions of the images. In Section 4.1 we hypothesize as to how this might be done.

Additionally, the research performed for this dissertation confirmed early speculation in the computer vision field that a coaxial camera rig might have advantages over a comparable stereo rig in terms of reducing the size and frequency of occlusions. In Section 4.2 we take a closer look at the coaxial camera vs. the stereo rig in occluded areas.

We used two numerical methods to solve the energy formulation. In Section 4.3 we compare the two and make some brief observations about the results from each.

Lastly we provide some suggestions for future work using motion-based correspondences.

4.1 Potential for Motion-Based Correspondences in Scene Flow

Scene flow is the estimation of the 3D scene motion field using a combination of disparity estimation and optical flow. There are two main approaches to scene flow:

coupled and decoupled [1]-[6]. In coupled approaches, the depth estimate and temporal tracking problems are solved simultaneously. In the decoupled approach depth from disparity is solved independently from the optical flow and then the two are combined.

Wedel et al. [3] report that decoupled methods of estimating scene flow are more effective than coupled methods. This result is presumably because the most accurate optical flow estimation uses variational methods whereas the most accurate disparity estimation methods use graphing techniques. In decoupled methods of estimating scene flow, disparity is first computed using intercamera correspondences. Optical flow is computed using intracamera temporal image pairs. The depth is then combined with the optical flow to estimate the 3D motion in the scene. Optical flow provides the XY displacement of points in the scene scaled by the depth estimate, and the Z motion in the scene comes from the change in the depth map over time.

However, decoupled methods do not take advantage of the additional information that comes from aligning the optical flow fields. Combining optical flow field derived correspondences with intercamera image-feature- or pixel-intensity-based correspondence produces a redundant set of correspondences based on different scene information (intensities and/or features vs. motion). Where the two sets of correspondences do not match, they provide insight into the error in the scene flow estimation as well as a consistency constraint in the optical flow computation. One could foreseeably use this optical flow consistency constraint to improve the estimation of the optical flow, thereby improving the overall accuracy of decoupled scene flow estimation.

4.2 Coaxial Camera Rig versus Multimodal Stereo Rig—Occlusions

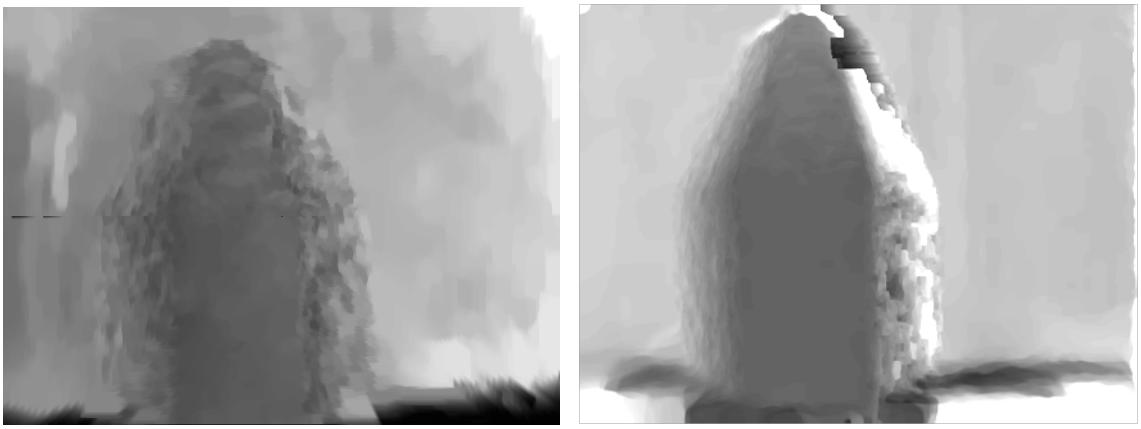
The scene flow accuracy is comparable in the nonoccluded areas of the scene, but there is a noticeable difference in the size of the occlusions between the coaxial camera rig and the multimodal stereo rig. Ma and Olsen [15] speculated that depth from zooming, the predecessor of the coaxial camera rig, would produce fewer occlusions than a similar binocular stereo rig, but because they were unable to reconstruct the center region of a depth from zooming image pair, they were unable to demonstrate this potential advantage. Doing a side-by-side image comparison between images taken with the coaxial camera rig vs. those taken with the stereo rig shows convincingly the advantages of the coaxial camera relative to the minimization of occlusions.

Figures 4.1, 4.2, and 4.3 show the graph-cuts-derived images for the three scenes, with a side-by-side comparison of the coaxial camera rig and the multimodal camera rig. The overall visual quality of the depth maps is similar between the coaxial camera and the stereo camera derived reconstructions, but the anomalies due to occlusions in the coaxial camera rig are dramatically smaller. This difference is most obvious in the fountain scene (Figure 4.1) on either side of the fountain. In the reconstruction from the coaxial image sequence (Figure 4.1 (a)) the fountain has the same general size and shape as in the source images, whereas in the reconstruction from the stereo rig, the occlusions on both sides of the fountain produce significant distortion of fountain width in the occluded areas.

We see similar issues with the stereo image in the upper right back-sloping surface of the flagstone image in Figure 4.2 (b), whereas the coaxial camera derived reconstruction does not exhibit this distortion. Lastly, in the flagstone plus alligator image (Figure 4.3

(a) (b)

Figure 4.1. Comparison of reconstructed depth maps from images taken with a coaxial camera rig vs. images taken with a multimodal stereo camera rig. Fountain scene. Graph cuts optimization. (a) Coaxial camera rig and (b) multimodal stereo camera rig.



(a) (b)

Figure 4.2. Comparison of reconstructed depth maps from images taken with a coaxial camera rig vs. images taken with a multimodal stereo camera rig. Flagstone scene. Graph cuts optimization. (a) Coaxial camera rig and (b) multimodal stereo camera rig.

(a) (b)

Figure 4.3. Comparison of reconstructed depth maps from images taken with a coaxial camera rig vs. images taken with a multimodal stereo camera rig. Flagstone with alligator scene. Graph cuts optimization. (a) Coaxial camera rig and (b) multimodal stereo camera rig.

(a) and (b)) the coaxial camera rig produces images that result in a substantially better reconstruction of the fine features of the alligator's mouth.

While not explicitly explored in this dissertation, occlusions and violations of the ordering constraint (points in the sensed image being in the same order as the corresponding points in the reference image) are connected. This connection suggests that image pairs from a coaxial camera rig would produce fewer violations of the ordering constraint than images taken with a traditional binocular stereo camera rig.

The reduction in occlusions as well as the potential of fewer violations of the ordering constraint suggests that a coaxial camera may have advantages over a binocular stereo camera rig in applications where occlusions are particularly problematic. Table 4.1 summarizes the differences between the coaxial camera rig and the stereo multimodal camera rig.

Table 4.1 Summary of differences between the coaxial and multimodal camera rigs.

	Coaxial camera Rig	Multimodal (stereo) Camera Rig
Camera Axes Alignment	Collinear	Parallel
Occlusions (Fountain Scene)	3-5 pixels	40-70 pixels
Minimum Working Distance (50% image overlap)	0 mm	190 mm
ICPM Flow Alignment Error - VM	< 0.01 pixels	0.09 - 0.12 pixels
ICPM Scene Flow Error - VM	3.1% - 3.9%	1.1% - 4.2%
ICPM Flow Alignment Error - GC	0.03 - 0.13 pixels	0.12 - 0.17 pixels
ICPM Scene Flow Error - GC	1.3% - 3.6%	1.6% - 2.2 %

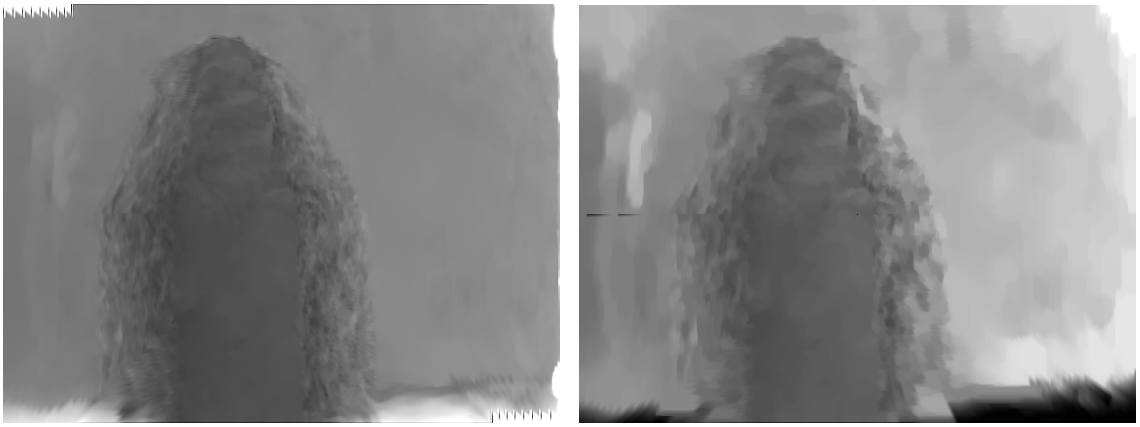
4.3 Variational Methods versus Graph Cuts

We explored two ways of solving the energy-minimization problem, variational methods and graph cuts. Variational methods form the basis of many if not most optical flow computations whereas graph cuts is the most widely used methodology for finding stereo correspondences.

In this dissertation, we solved similar energy-minimization problems for the same scenes using both methods. This two-solution approach produces an interesting comparison of the two techniques. Figures 4.4, 4.5, and 4.6 are a side-by-side comparison of the 3D reconstruction from images aligned using variational methods vs. 3D reconstruction from images aligned using graph cuts. Graph cuts produces the well-known "blocky" effect that is clearly visible in the graph cuts depth maps, but the overall reconstruction is comparable. Where visualization of the 3D structure is the objective, variational methods clearly have an advantage over graph cuts.

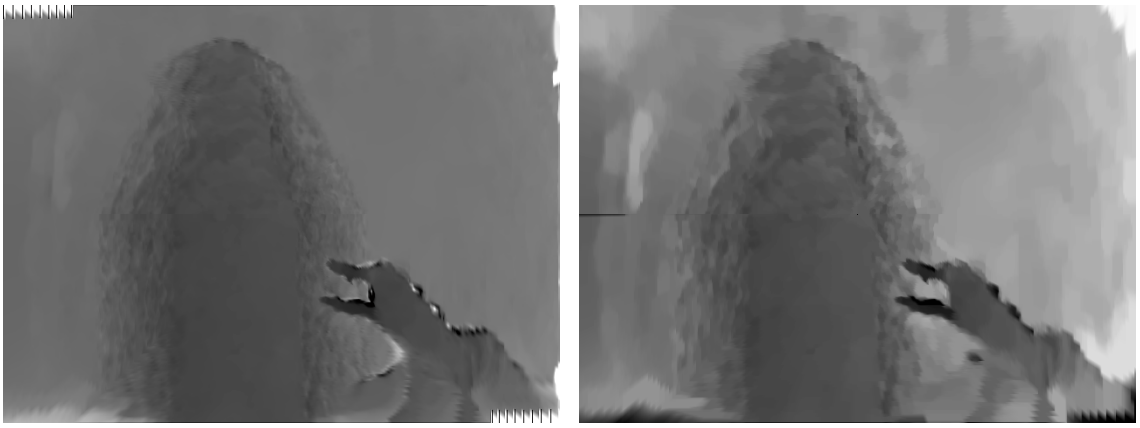
(a) (b)

Figure 4.4. Comparison of variational methods and graph cuts, coaxial camera rig, fountain scene. (a) Variational methods and (b) graph cuts.



(a) (b)

Figure 4.5. Comparison of variational methods and graph cuts, multimodal stereo camera rig, flagstone scene. (a) Variational methods and (b) graph cuts.



(a) (b)

Figure 4.6. Comparison of variational methods and graph cuts, multimodal stereo camera rig, flagstone plus alligator scene. (a) Variational methods and (b) graph cuts.

4.4 Future Work

As we have shown, motion-based correspondences can effectively be used to align images and produce realistic depth maps where image-feature- or pixel-intensity-based methods do not work. However, when image features or pixel intensities can be used, motion-based correspondences provide a redundant set of correspondences based on different information. As such, they provide an independent "opinion" on image alignment. The availability of this additional information suggests that combining motion-based correspondences with image-feature- or pixel-intensity-based correspondence-finding techniques might produce overall better results in a wide range of computer vision applications.

We have already discussed the possibility of using motion-based correspondences to improve decoupled scene flow estimation. In addition to scene flow, in the outer region of the coaxial camera where the radial disparity is larger than the disparity due to the projected motion, image-feature or pixel-intensity-based correspondences might produce better depth estimates. Combining the two methods for a coaxial camera could result in

improved overall results.

One of the more promising devices that motion-based correspondences enable is a 3D endoscope. For a 3D endoscope to be useful, reconstruction needs to be real-time. The second main area for future research would be to improve the computational efficiency such that 3D reconstruction could be done in real-time.

REFERENCES

[1] F. Huguet and F. Devernay, "A variational method for scene flow estimation from stereo sequences," INRIA, Paris, FR, Tech. Rep. 00166589v1, 2007.

[2] T. Basha, Y. Moses, and N. Kiryati, "Multi-view scene flow estimation: A view centered variational approach," *Int. J. Comput. Vis.*, vol. 101, pp. 6-21, April 2012.

[3] A. Wedel, *et al.*, "Stereoscopic scene flow computation for 3D motion understanding," *Int. J. Comput. Vis.*, vol. 95, pp. 29-51, October 2010.

[4] Y. Zhang and C. Kambhamettu, "On 3d scene flow and structure estimation," in *Computer Vision and Pattern Recognition*, Kauai, 2001, pp.778-785.

[5] S. Vedula, S. Baker, R. Rander, R. Collins, and T. Kanade, "Three dimensional scene flow," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, pp. 475-480, January 2005.

[6] R. Li and S. Sclaroff, "Multi-scale 3D scene flow from binocular stereo sequences," *Comput. Vis. Image Understanding*, vol. 110, pp. 75-90, April 2008.

[7] R. Kirby, "Ski speed determination system," U.S. Patent 7,564,477, July 21, 2009.

[8] R. Kirby, "Athletic response to using a real-time optical navigation feedback system during ski training," in *Int. Congr. Science Skiing IV*, St. Christof, Austria, pp. 304-309, 2007.

[9] R. Kirby, "Development of a real-time performance measurement and feedback system for alpine skiers," *Sports Technol.*, vol. 2, pp. 43-52, April 2009.

[10] R. Kirby and L. Karlöf, "Evaluating ski friction as a function of velocity using optical flow and hall effect sensors," in *Int. Congr. Science Skiing VI*, St. Chirstof, Austria, pp. 430-438, 2013.

[11] R. Kirby, "Understanding ski glide test data," in *Int. Cong. Science Skiing*, St. Christof, Austria, (in press).

[12] R. Kirby, J. Christian, G. Harmsen, A. Bennett, B. Markle, T. Laakso, *et al.*, "Multifunction snowpack measurement tool," US Patent Application 16/38,789, 2016.

[13] R. Kirby, "Three dimensional surface mapping system using optical flow," US Patent 8,860,930, September 14, 2014.

[14] R. Kirby, "Portable swing analyzer," US Patent 7,536,033, May 19, 2009.

[15] J. Ma and S. I. Olsen, "Depth from zooming," *J. Opt. Soc. Am. A* vol. 7, pp. 1883-1890, October 1990.

[16] R. Kirby and R. Whitaker, "Three dimensional coaxial endoscope," US Patent Application #62/374,998, August 15, 2016.

[17] R. Kirby and R. Whitaker, "3D reconstruction from images taken with a coaxial camera rig, " in *SPIE Optics + Photonics*, San Diego, 2016.

[18] J. Lavest, G. Rives, and M. Dhome, "Three dimensional reconstruction by zooming," *IEEE Trans. Robotics Autom.,* vol. 9, pp. 196-207, August 1993.

[19] J. Lavest, G. Reves, and M. Dhome, "Modeling an object of revolution by zooming," *IEEE Transactions on Robot. Autom.,* vol. 2, pp. 267-271, August 1995.

[20] N. Asada, M. Baba, and A. Oda, "Depth from blur by zooming," in *Vision Interface Ann. Conf.*, Ottawa, Canada, 2001.

[21] M. Baba, N. Asada, and T. Migita, "A thin lens based camera model for depth estimation from defocus and translation by zooming," in *15th Int. Conf. Vision Interface,* Calgary, Canada, 2002, pp. 274-281.

[22] H. Gao, J. Liu, Y. Yu, and Y. Li, "Distance measurement of zooming image for a mobile robot," *Int. J. Control, Autom. Syst.,* vol. 11, pp. 782-789, August 2013.

[23] Y. Zhang and K. Qi, "Snake-search algorithm for stereo vision reconstruction via monocular system," in *The 5th Ann. IEEE Conf. Cyber Technol. Autom. Control, Intell. Syst.*, Shenyang, China, 2015, pp. 497-502.

[24] T. Brox and J. Malik, "Large displacement optical flow desriptor matching in variational motion estimation," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 33, pp. 500-512, August 2010.

[25] R. Kirby and R. Whitaker, "A novel automated method for doing registration and 3D reconstruction from multi-modal RGB/IR image sequences," in *SPIE Optics + Photonics*, San Diego, 2016.

[26] T. Poggio, V. Torre, and C. Koch, "Computational vision and regularization theory," *Nature*, vol. 317, pp. 314-319, September 1985.

[27] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Vis.*, vol. 6, no. 6, pp. 721-741, January 1984.

[28] J. Marroquin, S. Mitter, and T. Poggio, "Probabilistic solution of ill-posed problems in computational vision," *J. Am. Stat. Assoc.*, vol. 82, pp. 76-89, March 1987.

[29] S. T. Barnard, "Stochastic stereo matching over scale," *Int. J. Comput. Vis.*, vol. 3, no. 1, pp. 17-32, May 1989.

[30] D. Marr and T. Poggio, "Cooperative computation of stereo disparity," *Science* vol. 194, pp. 283-287, October 1976.

[31] D. Scharstein, R. Szeliski, and R. Zabih, "A taxonomy and evaluation of dense two frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, pp. 7-42, April 2002.

[32] P.-J. Käck, "Robust stereo correspondence using graph cuts," M.S. thesis, Dept. Comput. Sci., KTH, Stockholm, Sweden, 2004.

[33] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, pp. 1222-1239, August 1998.

[34] S. Roy and I. Cox, "A maximum-flow formulation of the N-camera stereo correspondence problem," in *Int. Conf. Comput. Vis.*, Bombai, India, 1998, pp. 492-499.

[35] Y. Boykov and V. Kolmogorov, "Graph cuts in vision and graphics theories and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 79-96, September 2004.

[36] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intell.*, vol. 17, pp. 185-203, August 1981.

[37] A. Bruhn, J. Weickert, and C. Schnorr, "Lucas/Kande meets Horn/Shunck: Combining local and global optic flow methods," *Int. J. Comput. Vis.*, vol. 61, pp. 211-231, February 2005.

[38] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *8th Eur. Conf. Comput. Vis.*, 2004, pp. 25-36.

[39] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *Int. J. Compu. Vis.*, vol. 92, pp. 1-31, March 2010.

[40] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, pp. 359-374, July 2001.

[41] J.-Y. Bouguet. (2015-05-15). *Camera Calibration Toolbox for MATLAB* [software]. Available: http://www.vision.caltech.edu/bouguetj/calib_doc/

[42] Z. Zhang, "Flexible camera calibration by viewing a plane from unknown orientations," in *7th IEEE Int. Conf. Comput. Vis.*, Toronto, Canada, 1999, pp. 667-673.

[43] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, pp. 1330-1334, August 2000.

[44] L. Ford and D. Fulkerson, *Flows in Networks*. Princeton, NJ, USA: Princeton University Press, 1962.

[45] V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts?," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, pp. 147-159, June 2004.

[46] A. Delong and H. Osokin, "Fast approximate energy minimization with label costs," *Int. J. Comput. Vis.*, vol. 96, pp. 1-27, January 2012.

[47] O. Veksler and A. Delong. (2017-06-17). *GCoptimization* [software]. Available: http://www.csd.uwo.ca/faculty/olga/software.html

[48] P. Viola and W. M. Wells, "Alignment by maximization of mutual information," *Intl. J. Comput. Vis.*, vol. 24, pp. 137-154, September 1997.

[49] M. Yaman and S. Kalkan, "An iterative adaptive multi-modal stereo-vision method using mutual information," *J. Vis. Commun. Image Representation*, vol. 26, pp. 115-131, January 2015.

[50] G. Egnal, "Mutual information as a stereo correspondence measure," *Technical Report MS-CIS-00-20,* Computer and Information Science, University of Pennsylvania, Philadelphia, USA, 2000.

[51] C. Fookes, A. Lamanna, and M. Bennamoun, "A new stereo image matching technique using mutual information," in *Int. Conf. Comput., Graphics Imaging*, Honolulu, USA, 2001.

[52] C. Fookes, S. Maeder, S. Sridharan, and J. Cook, "Multi-spectral stereo image matching using mutual information," in *2nd Int. Symp. 3D Data Process., Vis., Transmission,* Thessaloniki, Greece, 2004, pp. 961-968.

[53] S. Krotosky and M. Trivedi, "Multimodal stereo image registration for pedestrian detection," in *IEEE Intell. Transp. Syst. Conf.*, September 2006, pp. 109–114.

[54] S. Krotosky and T. Mohan, "Registration of multimodal stereo images using disparity voting from correspondence windows," in *IEEE Conf. Advanced Video and Signal based Surveillance*, Sydney, Australia, 2006.

[55] S. Krotosky and M. Trivedi, "Mutual information based registration of multimodal stereo videos for person tracking," *Comput. Vis. Image Understanding*, vol. 106, pp. 270-287, May–June 2007.

[56] F. B. Campo, R. L. Ruiz, and A. D. Sappa, "Multimodal stereo vision system: 3D data extraction and algorithm evaluation," *IEEE J. Sel. Topics Signal Process.*, vol. 6, pp. 437-446, June 2012.

[57] A. Toraby and G. Bilodeau, "Local self-similarity as a dense stereo correspondence measure for themal visible video registration," in *2011 IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recognition Workshops*, Colorado Spring, CO, 2011, pp. 61-67.

[58] C. Aguilera, F. Barrera, F. Lumbreras, A. D. Sappa, and R. Toledo, "Multispectral image feature points," *Sensors*, vol. 12, pp. 12661-12672, September 2012.

[59] A. Verri and T. Poggio, "Motion field and optical flow: Qualitative properties," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, pp. 490-498, May 1989.